Doctoral Dissertation
博士論文

Automatic Metaphor Detection in Japanese and English
Using Machine Learning
(日本語及び英語における機械学習を用いたメタファーの自動検出)

Graduate School of Information Science and Technology
Hokkaido University

Mateusz Babieno

北海道大学　大学院情報科学研究科

バビェノ・マテウシュ

May 2022

学 位 論 文 内 容 の 要 旨

博士の専攻分野の名称　博士（情報科学）　氏名　Mateusz Babieno

学 位 論 文 題 名

Automatic Metaphor Detection in Japanese and English Using Machine Learning
（日本語及び英語における機械学習を用いたメタファーの自動検出）

比喩的な表現は日常会話の至る所で見られるにもかかわらず, 現在の自然言語処理の分野ではメタファーの処理が困難となっている. その原因の一つは, 人間と異なり, コンピュータには曖昧な表現を扱うための背景知識が不足していることである.

世界中の言語における比喩的な表現の裏に同じメタファーが見られるケースが多い. その一例として「良いは上・悪いは下」(GOOD IS UP・BAD IS DOWN) という概念メタファーが挙げられる. 「下品」, “a fallen country”, 「堕落した人格」, “I feel down lately”, 「結果が下がった」, “low standards”, 「最低なやつ」などは全てネガティブな感情を持ち, 「悪いは下」のメタファーに基づく表現である. そのような表現は言語を問わず日常的に使われ, そのメタファーの普遍性を示していると考えられる. したがって, メタファーをめぐる研究を, 1 つの対象言語に限定する必要はないと考えている. そこで, 本研究では対象言語を日本語及び英語の 2 ヶ国語とした.

本学位論文の第一章では, 自然言語処理の立場より考慮した比喩表現をめぐる研究の重要さを示し, 本研究による本研究分野への貢献について述べる. 全体を理解しやすくするために, 比喩的な言語について述べ, それに関わる術語の定義を挙げる. 比喩的な言語の種類を羅列し, その定義の曖昧さを指摘しながら, それぞれの特徴を説明する.

第二章では, メタファーに対する言語哲学・認知言語学・情報科学的なアプローチについて述べ, それぞれの優位性や制限を指摘する. 現在に至るまで, 各分野において提案されてきた本研究の着想となった関連研究を紹介する.

第三章では, 日本語を対象に行ってきた実験について述べる. メタファー・データとして比喩辞典の項目を, その反対であるリテラルなデータとしてウイキペディアの記事などを用いた. 評価データとして, それぞれのクラスに属する文が含まれる青空文庫の小説文章を用い, 以下に述べる 2 つの実験を行った. 1 つ目の実験では, メタファーをめぐる認知言語学の仮説に基づいた特徴量の 7 種類を使い, それぞれのパフォーマンスへの影響を比較した. 2 つ目の実験では, 各文を 32,231 次元の単語ベクトルで表し, 3 つの分類アルゴリズム, すなわち SVM, Naïve Bayes, Random Forest の性能を比較した.

機械学習を用いた二値分類では, 両方のクラスを代表する学習データを収集することは重要な第一歩である. ところが, リテラルであるはずのデータに比喩表現も少なからず入っていることを確認した. この原因はメタファーはいかなる使用域にも見られるためである. 人手でラベルを付与された大量のデータが必要だと考えられるが, 本研究の著者の知る限りでは, メタファー検出のために構築した公開されているそのような日本語のデータセットは実験を行なっていた当時に存在しなかった. その上, 提案手法のスコアを他の既存手法のスコアと客観的に比較することが不可能であった.

第四章では, 英語を対象に行ってきた実験について述べている. メタファー自動検出研究で現時点で存在する主要なデータセットを 3 つ使用し, 既存手法との比較を行った. その大量の公開されて

いるデータセット (VUAMC, TroFi, MOH-X) にはメタファー・非メタファーという人手で付与したラベルが用意されているため, 世界の研究者の既存手法のスコアと提案手法によるスコアを比較し, その有意性を客観的に確認した. 提案手法は既存アルゴリズムを基に, 理論上適切な改善を 3 種類加え, 最終的にモデルを 3 つ提案した. その 1 つ目はある単語のリテラルな意味を表すよう, その単語の辞書の定義を導入したモデルである. 2 つ目のモデルではそれに加え, ある単語が比喩的に使われているか否かを確認するために, コサイン類似度を使用し, その単語のリテラルな意味と文脈内の意味との差を計るモデルである. 3 つ目のモデルでは他の機械学習手法よりも良質な文ベクトルを出力することが確認されている Sentence-BERT を 1 つ目・2 つ目のモデルで使用した RoBERTa 言語モデルの代わりに使用するモデルである. 合計 13 回の評価実験を通し, そのうち 8 回では, 提案手法が既存手法の F 値を上回ることを確認した. その主な理由は辞書の定義を導入したことによるものであると考えられる.

第五章では, 上述の内容のまとめとそれに基づいた結論や今後の課題について述べる. 英語を対象言語にした本研究の後半において, 辞書の定義が表す単語の基本的な意味と文脈上の意味とのギャップは, その単語の比喩的な使用を示唆することが多いと確認した. それを参考に, 必要に応じて新しいデータセットを用意し, データ・アノテーションを加え, 深層学習に基づいた現在日本語でも利用可能になった最新言語モデルを用い, 日本語に対して実験を再度行うことを今後の課題とする.

Contents

Structure of the Dissertation

This dissertation is composed of five chapters.

In Chapter 1, I present the definitions of some of the terms associated with figurative language and metaphor, the main subjects of my research. Stressing on their conceptual closeness to one another, I briefly describe a number of related categories like *metaphor*, *simile*, *metonymy*, and *synecdoche*. Eventually, I argue that incorporating all of them into the scope of my research is appropriate from the perspective of computational approach to natural language understanding.

Chapter 2 provides an overview of the selected publications related to the current study. Although the presented works fall under various categories of scientific research: philosophy of language, cognitive linguistics, and natural language processing, all of them have become an invaluable source of inspiration.

Chapter 3 presents the part of my research devoted to figurative language identification within Japanese. I describe two sets of classification experiments conducted using Japanese datasets, aiming at detecting sentences comprising expressions with non-literal meanings.

In Chapter 4, I portray the architecture of MIss RoBERTa

WiLDe (Metaphor Identification using Masked Language Model with Wiktionary Lexical Definitions), a model for metaphor detection I have presented recently. I report on a set of experiments conducted utilizing some of the most popular datasets designed for metaphor detection in English. The results yielded by three variants of my newest method are compared to the scores achieved by a number of state-of-the-art baseline models.

Chapter 5 summarizes my findings and discusses future directions for my research.

Chapter 1

Introduction

1.1 Background

Figurative language in general and metaphor in particular allow us to speak more concisely, amusingly, and evocatively than with only literal expressions. Besides, such an endeavor would be destined to fail in the first place. This is because the figurative use of words is so ubiquitous in our language that it is very likely to be encountered in any randomly selected passage of the newspaper [111]. Metaphors are widely used in politics [71, 52, 53], psychotherapy [84, 95, 51], marketing [17], journalism [42, 81], and other domains deeming persuasion highly valuable. Unfortunately, metaphorical language remains difficult to process for computers despite the significant progress that has taken place in the field of NLP (natural language processing) over the last few years. This fact alone is a considerable reason to work on improving already existing algorithms as well as creating new ones designed to overcome this issue.

Machine translation is one of the NLP's subfields that still struggles with handling metaphorical expressions. Consider the following example of English-to-Japanese translation:

- English: *Yuki is so sweet.*
- Japanese: *Yuki wa totemo amai.*

While the input phrase sounds perfectly fine in English, its output translation is seen as awkward by Japanese native speakers. The adjective *sweet* can be, and often is, used figuratively in English in the sense of 'kind, gentle, or nice to other people', but that is not the case in Japanese. The adjective *amai* is often used metaphorically as a noun modifier, but if used to describe a person's personality traits, it conveys a very different meaning, specifically 'lenient, forgiving'. By analyzing the above example, it can be noticed that the input sentence is translated in its literal sense. The algorithm is seemingly not aware of the fact it has encountered a metaphor. As a result of this, it is taking *sweet* as a word belonging to the semantic field of TASTE rather than that of PERSONALITY FEATURES. As this is the output of the currently available version of Google Translate's engine[1], it should be clear that there is still much to do when it comes to improving the performance of the algorithms related to natural language understanding.

[1] https://translate.google.com/?sl=en&tl=ja&text=Yuki%20is%20so%20sweet&op=translate; last accessed on 1 March 2022.

1.2 Contribution

In this dissertation, I present the outcome of my research devoted to automatic metaphor identification using machine learning. It can be roughly divided into two main parts: research in metaphor detection in Japanese and research in metaphor detection in English.

In Chapter 3, I report on the two sets of experiments designed for sentence-level figurative language detection in Japanese. Since their presence in text poses similar problems from the perspective of automatic identification, in this part of the research I am searching not only for narrowly defined *classic* metaphors but also for related figures conveying non-literal meaning (i.e. metonymies, synecdoches, similes, idioms and proverbs). Because the borders separating non-literal and literal expressions are often vague, I argue that reformulating metaphor detection problem from the most often undertaken token-level detection to sentence-level detection results in a more flexible approach allowing for identifying diverse types of figurative language, including those linguistic units which — while encompassing larger portions of text — can be recognized as the manifestations of underlying conceptual feature mappings.

While most of the algorithms designed for metaphor identification in Japanese strongly depend on more or less complicated systems of rules, I propose a fully data-driven method, utilizing two kinds of input features. In the first experiment,

I take advantage of seven different seed word categories, selected based on the cognitive-linguistic view on metaphor. In the second trial, I use the bag-of-words model to represent the input sentences.

The methods I introduce in Chapter 3 were reported in the Language Sense on Computer workshop at 28th International Joint Conference on Artificial Intelligence, IJCAI 2019 and at the 16th International Conference of the Pacific Association for Computational Linguistics, PACLING 2019. They were first presented in [4] and [3], respectively.

In Chapter 4, I present the three variants of MIss RoBERTa WiLDe (Metaphor Identification using Masked Language Model with Wiktionary Lexical Definitions), a model built upon MelBERT (Metaphor-aware late interaction over BERT) introduced in 2021 by Choi et al. [16]. I found their work inspiring, not only because MelBERT outperforms current state-of-the-art in many cases, but also because their approach is well grounded in linguistics. While acknowledging the high quality of their work, I anticipated that applying some changes to their ideas would be appropriate from a theoretical perspective and could be beneficial to the algorithm's overall performance, making it more consistent conceptually.

Specifically, I argued that introducing the lexical definition of the target word in place of the target word itself is better-suited for finding the word's refined literal sense. I also estimated that using a different kind of sentence embedding representation should allow for achieving even better scores.

The results I present in Chapter 4 mostly confirm my intuitions. The model was first presented in the Applied Sciences journal article [5].

In this dissertation I use the contents of the three aforementioned works [4, 3, 5], modifying them where appropriate.

1.3 Terms of Figurative Language

In order to facilitate better understanding of this dissertation's topic, I provide the reader with a number of definitions describing some of the most important concepts recurring in the following chapters.

Figurative language

In the first sentence of this work's introduction, I have used the term *figurative language*. *A Glossary of Literary Terms* by Abrams & Harpham [1, p. 132] defines it as:

> a conspicuous departure from what competent users of a language apprehend as the standard meaning of words, or else the standard order of words, in order to achieve some special meaning or effect.

While figurative language is often deemed as belonging to the domain of poetry, Abrams & Harpham [1, p. 132] explain that:

> figures are sometimes described as primarily poetic, but they are integral to the functioning of language and indispensable to all modes of discourse.

This description further differentiates between two classes of figurative language: *figures of thought* (or *tropes*) and *figures of speech* (or *schemes*), in which, respectively:

1. Words or phrases are used in a way that effects a conspicuous change in what we take to be their standard meaning. The standard meaning, as opposed to its meaning in the figurative use, is called the *literal meaning*.
2. The departure from standard usage is not primarily in the meaning of the words but in the order or syntactical pattern of the words.

Subsequently, the editors specify a number of figures of thought, among them: *simile*, *metaphor*, *metonymy*, *synecdoche* and *personification*. According to this view, figures of speech comprise such rhetorical devices as: *anaphora*, *anastrophe*, *apostrophe*, *chiasmus*, *paralipsis*, *rhetorical question* and *zeugma*.

However, as Abrams & Harpham admit [1, p. 133], this "distinction is not a sharp one, nor do all critics agree on its application". Indeed, other sources recognize only *figures of speech* as a broader category encompassing all of the elements belonging to both classes listed above. For example, according to Britannica Concise Encyclopedia [11]:

> common figures of speech include simile, metaphor, personification, hyperbole, irony, alliteration, onomatopoeia, and puns.

The same source defines *figure of speech* as a:

> form of expression used to convey meaning or heighten effect, often by comparing or identifying one thing with another that has a meaning or connotation familiar to the reader or listener. An integral part of language, figures of speech are found in oral literatures as well as in polished poetry and prose and in everyday speech.

Regardless of the differences in taxonomic nuances, both sources agree on that figures play an important role in the ordinary language. The reason for their high frequency is explained by Gibbs in [31, p. 253]:

> I shall argue that a major reason why people use different tropes so frequently in everyday speech and writing is that human cognition is fundamentally shaped by various processes of figuration. (...) Speakers can't help but employ tropes in everyday conversation because they conceptualize much of their experience through the figurative schemes of metaphor, metonymy, irony, and so on. Listeners find tropes easy to understand precisely because much of their thinking is constrained by figurative processes.

1.3.1 Metaphor

Abrams & Harpham [1, p. 133] write that in metaphor:

> a word or expression that in literal usage denotes one kind of thing is applied to a distinctly different kind of thing, without asserting a comparison.

Using a line from one of the poems by Robert Burns: "O my love is like a red, red rose", they show that metaphor can function as a condensed simile:

> if Burns had said "O my love is a red, red rose" he would have uttered, technically speaking, a metaphor instead of simile.

Metaphorical expressions of syntax *A is B* are often used as examples in dictionaries, but there is a great variety of shapes that metaphors can take. Some of them can be seen in the definition of metaphor available in Britannica Concise Encyclopedia [11]:

> Figure of speech in which a word or phrase denoting one kind of object or action is used in place of another to suggest a likeness or analogy between them (as in "the ship plows the seas" or "a volley of oaths"). A metaphor is an implied comparison (as in "a marble brow"), in contrast to the explicit comparison of the simile ("a brow white as marble").

Poetic style of metaphorical expressions provided so far may wrongly suggest that metaphor belongs to literature and cannot be found in ordinary language. To the contrary — as convincingly claimed by cognitive linguists — most people use metaphors regularly without realizing they do. In [101, p. 51], Steen argues that metaphor is a:

> part of common, everyday language, as is attested by the many metaphorical forms (words, phrases, morphemes, and even grammatical constructions) that are entirely conventional.

This intuition is supported by Gibbs [32, p. 3], who reports that:

> there is now a huge body of empirical work from many academic disciplines that clearly demonstrates the ubiquity of metaphor in both everyday and specialized language.

Cognitive linguistics propose a view in which metaphor is identified with the psychological process underlying the production of metaphorical expressions rather than with such expressions themselves. This process consists in a cross-domain mapping of features[2]. For example, in TIME IS MONEY metaphor, the properties primarily ascribed to the concept of money are

[2]There are several competing variants of this theory, e.g. the two-domain approach, the many-space approach, the class-inclusion approach, the career of metaphor approach, etc. For details, cf. [97, pp.48-57].

transferred onto the concept of time. This results in production of expressions like: *I've invested a lot of time in this research* or *This solution will save you years of work.* To quote Gibbs again [32, p. 3]:

> Metaphor is not simply an ornamental aspect of language, but a fundamental scheme by which people conceptualize the world and their own activities.

1.3.2 Metonymy

Some authors suggest that in the hierarchy of importance to human's thought, it is metonymy that should be given precedence over metaphor. Consider the following passage from the introduction to the book by Panther & Radden [80, p. 1]:

> Eighteen years after Lakoff and Johnson's (1980) seminal work on the role of metaphor in conceptualization, which sparked a vast amount of research in cognitive linguistics, it has become increasingly apparent that metonymy is a cognitive phenomenon that may be even more fundamental than metaphor.

In Encyclopaedia Britannica[3], metonymy is described as a:

> figure of speech in which the name of an object or concept is replaced with a word closely related to or suggested by the original, as "crown" to mean "king" ("The power of the crown was mortally weakened") or an author for his works ("I'm studying Shakespeare").

The same article notes on the similarity between metonymy and synecdoche:

[3] https://www.britannica.com/art/metonymy; last accessed on 1 March 2022.

> Metonymy is closely related to synecdoche, the naming of a part for the whole or a whole for the part, and is a common poetic device.

Although this remark may suggest primarily poetic nature of metonymy, it is widely used in everyday language. For example, *the article* stands for *the article's author* in "the article claims that...". Similarly, in "White House denied the report", *White House* denotes *the President of the United States*.

Abrams & Harpham [1, p. 134] emphasize the experiential roots of the trope in question, writing that in metonymy:

> the literal term for one thing is applied to another with which it has become closely associated because of a recurrent relation in common experience.

Regarding the relation between metaphor and metonymy, cognitive linguistics offers the view on which both phenomena involve a mapping of semantic features. Whereas metaphor is seen as the mapping across the distinct conceptual domains, metonymy consists in the "mapping within the same domain" [110, p.114]. While metaphor is based on imagined resemblance between the interacting concepts (e.g. *time* and *money*), metonymy has to do with their contiguity (e.g. *king* and *crown*).

1.3.3 Synecdoche

As mentioned above, synecdoche is conceptually very similar to metonymy, and some accounts even treat the former as a

subspecies of the latter. In Encyclopaedia Britannica[4] it is defined as:

> figure of speech in which a part represents the whole, as in the expression "hired hands" for workmen or, less commonly, the whole represents a part, as in the use of the word "society" to mean high society. Closely related to metonymy — the replacement of a word by one closely related to the original — synecdoche is an important poetic device for creating vivid imagery. An example is Samuel Taylor Coleridge's line in "The Rime of the Ancient Mariner", "The western wave was all aflame", in which "wave" substitutes for "sea".

In other words, synecdoche can be viewed as a means for "generalization and particularization of senses" [97, p. 147].

That synecdoche is not merely a "poetic device" can be proved by the following examples from everyday language: in "Let's go and have a bite", *a bite* stands for *a meal*; in "I don't have a penny left", *a penny* stands for *money*.

1.3.4 Personification

In Encyclopaedia Britannica[5] personification is defined as a:

> figure of speech in which human characteristics are attributed to an abstract quality, animal, or inanimate object.

Abrams & Harpham [1, p. 135] quote Milton's example of personification used in his famous *Paradise Lost*:

> Sky lowered, and muttering thunder, some sad drops
> Wept at completing of the mortal sin.

[4] https://www.britannica.com/art/synecdoche; last accessed on 1 March 2022.
[5] https://www.britannica.com/art/personification; last accessed on 1 March 2022.

In the lines above, inanimate *thunder* and *rain* are described with the words applied primarily to human beings. Here, the particular combination of source and target domains reveals the metaphorical nature of the analyzed instances of personification. In other cases it can be used metonymically, as in the example of *the article claims that...* mentioned earlier.

The figures described so far can be realized as single words (e.g.: metaphor in *They* ***attacked*** *my argument*; metonymy in *This land belongs to the* ***crown***; synecdoche in *Give us this day our daily* ***bread***; personification in *Tokyo never* ***sleeps***).

There are, however, figurative units, that are inherently composed of more than a single word. One of those is already mentioned simile.

1.3.5 Simile

Encyclopaedia Britannica[6] defines simile as a:

> figure of speech involving a comparison between two unlike entities. In the simile, unlike the metaphor, the resemblance is explicitly indicated by the words "like" or "as". The common heritage of similes in everyday speech usually reflects simple comparisons based on the natural world or familiar domestic objects, as in "He eats like a bird", "He is as smart as a whip", or "He is as slow as molasses". In some cases the original aptness of the comparison is lost, as in the expression "dead as a doornail".

Japanese counterparts of English simile indicators listed in the above definition would be: *yō*, *mitai*, *gotoki*, etc.

[6] https://www.britannica.com/art/simile; last accessed on 1 March 2022.

Similes have a specific structure, which includes *comparandum*, *comparatum* and *tertium comparationis* [22, p. 44]. In the following example:

- *Her eyes are green as emeralds*,

eyes correspond to comparandum, *emeralds* to comparatum, and *green* to tertium comparationis.

It should be noted, that tertium comparationis is often omitted, which may increase the level of perceived figurativeness (e.g. *Her eyes are like emeralds*).

1.3.6 Idiom

Oxford Dictionary of English [103] defines idiom as:

> a group of words established by usage as having meaning not deducible from those of the individual words (e.g. *over the moon*, *see the light*).

In [22, pp. 39-44], Dobrovol'skij & Piirainen give an elaborate description of idioms, listing their characteristics, which include:

- Idiomacity understood as a semantic reinterpretation and/or opacity,
- Stability understood as frozenness or lack of combinatorial freedom of a certain expression,
- Being a multiword unit.

English *piece of cake* 'something very easy' and Japanese *kubi o nagaku suru* 'to eagerly look forward to something' (lit. 'to lengthen ones neck') are examples of idioms.

1.3.7 Proverb

Another category belonging to figurative language is that of proverbs. Oxford Dictionary of English [103] defines a proverb as "a short, well-known pithy saying, stating a general truth or piece of advice". Concise Encyclopaedia Brittanica [11] adds that proverbs:

> are part of every spoken language and folk literature, originating in oral tradition. Often a proverb is found with variations in many different parts of the world.

The relation connecting proverbs with idioms is well analyzed by Dobrovol'skij & Piirainen [22, pp. 49–53]. Like idioms, proverbs are intrinsically multi-word units. In many languages proverbs can be differentiated from idioms by their syntactic structure: while most idioms are phrases, proverbs often function as independent sentences. One example of a proverb provided by the authors is the following:

> *every dog has its day* 'even the most unimportant person has a time in his/her life when he/she is successful and being noticed'.

Although, as mentioned, proverbs usually take shape of independent sentences, sometimes they can be embedded within the longer sequences as well: *She believes that she has a chance, because every dog has its day.*

1.4 Scope of Investigation

Concepts related to figurative language are similar in many respects, which makes the borders separating one from another quite vague. For example, Abrams & Harpham [1, p. 133] note that "metonymy and synecdoche are sometimes categorized as species of metaphor". Moreover, the same expression can belong to multiple categories at the same time[7]. This issue is well portrayed by Schofer & Rice [90]:

> The very titles of recent works in the area suggest the muted struggle for the understanding and conquest of the terms: *Synecdoques*, in which Tzvetan Todorov (1970) defines metaphor as a double synecdoche; *Metonymie et metaphore*, where Albert Henry (1971) defines metaphor as a double metonymy; *Semantique de la metaphore et de la metonymie*, in which Michel Le Guern (1975) subsumes synecdoche within metonymy; and *La Metaphore vive*, where Paul Ricoeur (1975) reestablishes metaphor in its privileged position.

As signalized in Section 1.3.2, from the perspective proposed by cognitive linguistics, the demarcation line between metaphor and metonymy should be set based on the number of conceptual domains that are used in the process of feature-mapping. However, whether said process takes place within one (metonymy) or multiple (metaphor) domains cannot be always established unequivocally. Writing about difficulties in distinguishing metonymy from metaphor, Barcelona [7, p. 232] enumerates the three major ones:

[7] For example, in the study on so called *metaphtonymy* [34], the author analyzes examples of expressions exhibiting properties of both metaphor and metonymy simultaneously.

- Cognitive domains often have fuzzy boundaries so that it is not always easy to know if the source and the target domains are or are not in the same superordinate domain.
- A linguistic expression may often be interpreted, on the *sole* basis of context, background knowledge, or the purpose of the interpreter, as metaphorical, or as metonymic.
- Metaphor and metonymy very often interact in intricate patterns, which complicates their distinction.

This echoes in the analysis presented by Steen et al. [100, p. 79], who admit that "both metaphor and metonymy (...) can be present at the same time" and that "it is possible to see a conceptual or semantic relation as either or both metaphoric and metonymic".

Considerations on the nature of metaphor have a long tradition that dates back at least to antiquity. In reference to the Aristotelian definition of metaphor from *Poetics*[8], Turbayne [105, p. 11] calls the reader's attention to its capacity:

> Notice how wide Aristotle's definition is. Metaphor comprehends all those figures that some distinguish as: synechdoche (...); metonymy (...); catachresis (...); and metaphor (...).

and concludes that:

> [To Aristotle - M.B.] metaphor is logically indistinguishable from trope, the use of a word or phrase in a sense other than that which is proper to it.

[8] "Metaphor consists in giving the thing a name that belongs to something else; the transference being either from genus to species, or from species to genus, or from species to species, or on grounds of analogy" (cf. https://www.gutenberg.org/files/6763/6763-h/6763-h.htm; last accessed on 1 March 2022). Additionally, in his *Rethoric*, Aristotle states that "simile also is a metaphor", (cf. https://kairos.technorhetoric.net/stasis/2017/honeycutt/aristotle/rhet3-4.html; last accessed on 1 March 2022).

Such perspective can be identified with the broader view on metaphor, as described by Fogelin [29, p. 30]:

> Viewed narrowly, a metaphor is one trope among others — including similes, hyperbole, irony, metonymy, synecdoche, etc., etc. It is also possible to take 'metaphor' and 'metaphorical' in a more generic sense which is more or less coextensive with 'figure' and 'figurative'.

This is confirmed by the notation available in Merriam Webster[9], which states that — in a broad sense — the term *metaphor* can be used interchangeably with *figurative language*. Whereas Turbayne commends such an approach to defining a metaphor (he in fact advocates for widening it even more, cf. [105, p. 12]), Fogelin criticizes it, arguing that "an adequate theory of figurative language ought to provide ways of distinguishing various tropes from one another" [29, p. 32].

The goal set in the first part of this research introduced in Chapter 3 was to detect the instances of figurative language use, irrespective of their narrowly defined taxonomic affiliations. At the stage of labeling the testing-data, the annotators were asked to "decide whether any figurative expression (metaphor, simile, etc.) is used in the (...) sentence". The rationale behind this decision was that any type of the input whose intended meaning is to be interpreted as involving a departure from the most basic literal senses of its constituents may hinder the performance of Natural-Language-Understanding-based tools (e.g. machine translation engines).

[9] https://www.merriam-webster.com/dictionary/metaphor; last accessed on 1 March 2022.

From such perspective, it is not that much relevant which of the figurative mechanisms underlies a non-basic meaning conveyed by the input text. Both *crown* used in place of *king* and *down* used in place of *sad* may become a cause of an erroneous translation. It follows that not only metaphors, but also metonymies, synecdoches, similes, idioms, proverbs, and alike: all of them have become the targets of identification.

In the latter part of the research presented in Chapter 4 I was using the already labeled datasets. Their major source — VUA Metaphor Corpus (VUAMC)[10] described in details in Section 4.4 — was annotated with MIPVU[11], which distinguishes between metaphor and metonymy[12]. In VUAMC, only the words representing the former category are provided with labels. Consequently, in the experiments I report on in Chapter 4, the identification target is narrower than in those described in Chapter 3. However, as metonymy also involves a detachment from the word's most basic literal sense, the design of my proposed model should allow for detecting the words used metonymically as well.

The authors of MIPVU seem to consider the meanings conveyed by some of the expressions based on metonymy as the "extensions of the basic senses" which prefer literal interpretation [100, p. 80]. This might be due to the fact, that the

[10] http://www.vismet.org/metcor/documentation/home.html; last accessed on 1 March 2022.

[11] The upgraded variant of the Metaphor Identification Procedure (MIP; its outline is presented in Section 4.2.1).

[12] On the other hand, MIPVU treats similes and personifications as the (potential) forms of metaphor [100, pp. 21, 40–41, 57–58, 92–95, 101–106]. The complete guidelines for annotation are presented in a book-length description by Steen et al. [100].

term *literal* — as most of the terms analyzed in this chapter — is not free from ambiguity. In his paper [54], Lakoff elaborates on the multiple meanings of *literal*. The one I adopt in this research is similar to that used by the author in his technical discussions [54, p. 3]. Specifically, I consider *literal* as:

> *directly meaningful, without the intervention of any mechanism of indirect understanding such as metaphor or metonymy.*

In the Oxford Dictionary of English [103], *figurative* is defined as: "departing from a literal use of words; metaphorical". As the terms *non-literal*, *figurative*, and *metaphorical* are often used to denote the same meaning, in this dissertation I use them interchangeably.

Chapter 2

Research in Metaphor

In this chapter, I briefly introduce a number of works related to metaphor and figurative language that — in one way or another and to various degrees — influenced the final shape of this dissertation. Metaphor has become a subject of studies in many fields of academic research. The works I present below can be roughly classified as belonging to the following three: philosophy of language, cognitive linguistics and natural language processing.

2.1 Philosophy of Language

In his already classic paper from 1954 [9], American philosopher Max Black presents his considerations regarding metaphor. Here, I will touch upon the most important of his remarks.

Black proposes three different approaches in conceptualizing metaphor and labels them as respectively:

1. Substitution view,
2. Comparison view,
3. Interaction view.

On substitution view, a given metaphorical expression is used in place of a literal one, which conveys the same meaning. This account implies that the meaning of any metaphorical expression can be expressed with some literal counterpart.

On this view, one justified reason for speaking metaphorically is to fill the lexical gaps produced by the entities not yet named in a certain language. Such a case can be exemplified with the *leg* of an angle, the term known from mathematics. Other than that, writers use metaphors to simply amuse their readers, providing them with entertaining linguistic riddles. On this view, when not used to expand the vocabulary, metaphor is considered as a mere linguistic decoration.

Comparison view treats metaphor as a condensed, elliptic variant of simile. Black uses the example of *Richard is a lion*, which — on comparison view — is a shortened version of *Richard is like a lion* (in being brave). The author notices that this view does not explain the reasons for using only some of the features ascribable to the compared entities. Lions can be not only brave (can they be called *brave* literally is yet another issue), but also big, fast, hairy, scary, hungry, etc. Out of many possible candidates, why is it *a lion* that becomes employed by the metaphor, is another question that emerges when adopting comparison view on metaphor.

Black recommends interaction view as the most adequate perspective on metaphor. This view consists in metaphor using two different concepts (in the example above, the one of *a man* represented by *Richard* and the other of *a lion*), which

become simultaneously active in the reader's mind, and interact, producing a new meaning. These concepts are said to be based not on the lexical norms to be found in a standard dictionary, but rather on what Black calls the system of associated commonplaces, which are shared by all members of a given cultural community. On this view, in the process of constructing a new metaphorical meaning, the objective definition of *a lion* is less relevant than a set of associations connected to the concept of *a lion* in the mind of an average member of a given society. Simultaneously, another set of associations related to *a man* is evoked. While the traits of *a lion* that are also ascribable to *a man* become rendered as prominent, the leftovers lose their importance and become discarded.

Arguably, the main point of the interaction view proposed by Black consists in the idea that metaphor produces a new meaning that is different from a given expression's basic meaning. This perspective can be contrasted with Donald Davidson's account on metaphor [19], the overview of which I present below.

Davidson claims that metaphorical expressions do not hold any special meanings differing from the literal ones. He argues that they are often false in a logical sense, and — in this regard — they can be contrasted with similes, which are always true, since everything is similar to everything else, at least to a certain degree.

Accordingly, *a lion* from the sentence *Richard is a lion* does not mean anything else than *a lion* in any other context. Unless it is not a wild cat, specimen of *Panthera leo*, denoted by *Richard*, the sentence in question should be classified as logically false.

Although Davidson mostly disagrees with Black, he admits that a metaphor can be conceptualized as a filter allowing the reader to notice something about the outer world, that he might not have been fully aware of. However, this effect of encountering a metaphor should not be conflated with some special meaning that words allegedly hold in secret. To Davidson — in this respect — metaphors are similar to pieces of music or photographs: they are what they are, and nothing else. It is true that they can induce some special sensations by bringing some analogies and similarities to the receiver's attention, by intimating something surprising, but it is not because of some secret set of meanings concealed by the words. According to Davidson, there are no such secret meanings.

Since hidden metaphorical meanings do not exist, what is often called a metaphor's literal paraphrase should not be viewed as an accurate and objective explanation of the meaning. It is rather a description of interpreter's feelings accompanying him in his encounter with a metaphor. Because [19, p. 31]:

> understanding a metaphor is as much a creative endeavour as making a metaphor, and as little guided by rules,

no such paraphrase is ever objective — it may differ from person to person.

2.2 Cognitive Linguistics

In 1980, George Lakoff and Mark Johnson published their "Metaphors We Live By" [56], a book that has become one of the most frequently cited publications in the history of social sciences[1]. According to some authors, its issue marks the beginnings of cognitive linguistics as an autonomous branch of research [59, p. 140]. Lakoff & Johnson proposed a framework for metaphor analysis, which is now known as Conceptual Metaphor Theory (CMT). Below I describe some of its main assumptions.

According to CMT, a metaphor is a phenomenon belonging to the realm of thought and mental representation, rather than to the language on its own. What is colloquially labeled a metaphor — some metaphorical expression encountered in an utterance — is a manifestation of underlying cognitive process. In consequence, the term *metaphor* should be equated with this process rather than with a linguistic entity. In one of their early papers [55, p. 486], Lakoff & Johnson write that:

> Metaphorical concepts provide ways of understanding one kind of experience in terms of another kind of experience. Typically this involves understanding less concrete experiences in terms of more concrete and more highly structured experiences.

[1] According to the analysis from 2014 [35], #20 of all time. By the time of submitting this dissertation (March 2022), the number of citations on Google Scholar has reached almost 77,000.

Speaking about something abstract, we often use terms primarily belonging to some other concept, which we are more familiar with and which we understand better. A system of ideas related to a more abstract concept is called a *target domain*. Corresponding system of more concrete ideas used to describe the target concept is called a *source domain*. The mechanism of cross-domain *mapping* of features between distinct concepts lies at the core of Conceptual Metaphor Theory.

In [57, p. 195], Lakoff & Johnson argue that human conceptual system "is fundamentally metaphorical in character". They propose three types of metaphorical concepts, "which are realized by a vast number of linguistic expressions". These are:

1. Orientational metaphors: GOOD IS UP (*We hit a peak last year, but it's been going downhill ever since*), CONTROL IS UP (*he is under my control*), etc.
2. Ontological metaphors: THE MIND IS A CONTAINER (*He's empty-headed*), VITALITY IS A SUBSTANCE (*There's no life in him anymore since his accident*), etc.
3. Structural metaphors: UNDERSTANDING IS SEEING (*I see what you're saying*; *Now I've got the whole picture*), LIFE IS A GAMBLING GAME (*I've got an ace up my sleeve*), etc.

In [49, p. 57], Kasanuki provides a number of example sentences based on LOVE IS A JOURNEY structural metaphor[2]:

1. *Kono sannenkan, futari wa tomo ni ayunde kita.* ('For the last three years, the two have walked together.')

[2]The examples in English are cited from [56, p. 44-45]; translation of the Japanese items by myself.

2. *Donna shōgai ga atte mo, futari de norikoete ikō.* ('Whatever the obstacles are going to be, let the two of us overcome [lit. 'go over'] them')
3. We'll just have to go our separate ways.
4. Our marriage is on the rocks.
5. We've gotten of the track.

Through experiences, we notice similarities between various concepts, even those, which are dissimilar at first appearance. Kasanuki [49, pp. 57-58] explains that sentences listed above are based on the insight, that the concept of romantic love displays some common features with that of a journey.

For example, departure can be viewed as similar to the beginning of the romantic relationship; analogously, time used for a trip can be compared to a life spent together with the lover. Lakoff and Johnson point out [56, p. 45] that there are different types of journeys (car trip, train trip, sea voyage, etc.) and there is no single consistent *image*, which could encompass all of the expressions based on LOVE IS A JOURNEY metaphor.

It can be argued, that using non-obvious links between different conceptual domains increases the unexpectedness level of a given expression, making it *more metaphorical* in the addressee's perception.

Conceptual Metaphor Theory has influenced countless research papers, including some studies on Japanese figurative language. For example, Hiraga [44] analyzes the conceptual metaphor underlying such Japanese words as *bushidō*, *sadō*,

kadō, *shodō*, *kyudō*, *kendō* etc. As can be noticed, all of the compound words just listed include the morpheme *dō* 'way, road', which suggests that they are semantically interrelated.

Here it should be noted that Japanese is a multi-systemic language whose lexicon comprises three types of vocabularies, excluding hybrid words [48, pp. 12-13]. These three types are:

1. *Wago* (indigenous, native Japanese words; usually written with Hiragana or Hiragana and Kanji),
2. *Kango* (Sino-Japanese words originating in Chinese; usually written with Kanji),
3. *Gairaigo* (vocabulary originating in foreign languages other than Chinese – mostly English; usually written with Katakana).

The words including logographic character 道 *dō* listed above belong to the Sino-Japanese lexical subsystem *Kango*. The morpheme *dō* is suffixed to another morpheme (or multiple morphemes as *bushi* in *bushidō*), which results in formation of the words denoting classical disciplines (*sadō* 'Japanese tea ceremony', lit. 'tea' & 'road'; *kadō* 'ikebana', lit. 'flower' & 'road'; *shodō* 'Japanese calligraphy', lit. 'pen' & 'road'), martial arts (*kyudō* 'Japanese archery', lit. 'bow' & 'road'; *kendō* 'kendo', lit. 'sword' & 'road'), and philosophical systems (*bushidō* 'bushido', lit. 'warrior' & 'road'). This kind of word structure allows for immediate identification of the underlying LEARNING IS A JOURNEY (or LEARNING IS FOLLOWING A PATH) metaphor.

2.3 Natural Language Processing

History, methods, and goals of the computational approaches to metaphor are comprehensively described in [108]. In the following two subsections, I briefly introduce a number of papers focusing on figurative language detection in Japanese and English, respectively.

2.3.1 Japanese as the Target Language

In [68], Masui et al. undertake a simile identification task. In order to distinguish similes from literal and nonsensical comparisons sharing the syntactical pattern *B no yōna A* 'A like B', the authors adopt two kinds of parameters: *salience gap* (*kengensei rakusa*) and *novelty* (*igaisei*)[3].

The choice of the features can be justified with the following reasoning: in order to explain a concept *A* with a concept *B*, the salience of a predicate *X*, which is shared by both concepts, needs to be higher in a concept *B*; in order to leave a strong impression in the addressee, the juxtaposition of the concepts needs to be unexpected and original as well.

Salience gap can be illustrated with the example of simile provided by Ortony in [77, p. 346]: *Encyclopedias are like gold mines. Source of highly valuable resources* is arguably the most salient property of the *gold mine* concept, whereas in *encyclopedia* — although present — perhaps it occupies a lower position in the corresponding hierarchy.

[3]These are the English counterparts provided by the authors in the abstract. Arguably, *level of unexpectedness* might be a better translation of *igaisei* than *novelty*.

If compared concepts share no common predicates, such expression does not produce a salience gap and should be deemed nonsensical. Masui et. al portray this with the example of *Tanisoko no yōna kuruma* 'A car like a valley' [68, p. 74]. Salience gap does not occur in literal comparisons either, since the juxtaposed entities share too many properties. This can be illustrated with the examples: *Encyclopedias are like dictionaries* and *Gold mines are like oil wells* provided by Ortony in [77, p. 350].

As mentioned, the second of the metaphoricity parameters adopted by Masui et al. is novelty. The authors argue that high frequency in a corpus indicates a literal usage of a given word-pair. They explain that the comparison *supōtsukā no yōna kuruma* 'a car like a sports car' should not be considered novel, because it is encountered in Japanese too frequently.

The objects of said comparison are connected by hyponymy relation (*supōtsukā* 'a sports car' is a hyponym of *kuruma* 'a car'). Whereas the concept of a *sports car* comprises such predicates as *fast*, *cool*, *excessive fuel consumption* at the very top of its salience list, the same cannot be said about the concept of an average prototypical *car*, where the same predicates occupy lower positions. Although comparing these two entities triggers the emergence of a salience gap, the output expression should not be considered a simile. Since one term is a hypernym of the other, the concepts they represent share too many properties for their relationship to be considered metaphorical.

The authors do not elaborate on the correlation between metaphoricity and hyponymy, but it seems rather safe to assume that most of similar comparisons are indeed non-figurative. The examples: *a fruit like a pineapple*, *a bird like a penguin* or *Sapporo no yōna machi* 'a city like Sapporo' convey rather literal senses, which seems to confirm this intuition.

Following such theoretical considerations, the authors construct two knowledge bases, utilizing the corpus composed of approximately 110,000 newspaper articles. The first knowledge base comprises lists of the target nouns along with their properties and corresponding salience level expressed in percentage. The properties of a given concept are collected automatically. Having processed the corpus with a morphological parser[4], the authors obtain the word-pairs sharing the grammatical pattern of *adjectival modifier—noun* (e.g. *akai hana* 'red flower'). Most frequent modifiers of a given noun are regarded as its concept's most salient properties. The second knowledge base includes lists of noun-pairs along with their absolute frequencies in the corpus and their intrasentential co-occurrence frequencies. The pairs are composed of nouns appearing together in the span of one sentence.

In order to test their method, authors prepare two datasets comprising 100 expressions of shape *B no yōna A*, which come from the two corpora — the one which was used to create the already mentioned knowledge bases (70 expressions) and the other one comprising unrelated data (30 expressions).

[4] https://archive.ph/Rpgge; last accessed on 1 March 2022.

The analysis of samples appearing in the first dataset allows the authors to set up a number of threshold values for salience gap and novelty, which are used to establish the borders separating the three categories: similes, literal comparisons, and nonsensical expressions.

Having prepared the thresholds, the authors use a number of mathematical formulas in order to classify a given expression as belonging to one of the aforementioned categories. The experiments conducted on the data comprised by the two subsets mentioned above led authors to yield the following results: precision 83.3%, recall 52.1% (Dataset 1); precision 72.7%, recall 61.5% (Dataset 2).

In [74], Nekomoto et al. propose a system of heuristics for metaphor identification. They focus on the phrases of syntax: *A wa B* 'A is B' and *B no A* 'B's A', where both *A* and *B* are limited to nouns. With such restrictions, authors try to predict, whether the relation between *A* and *B* is figuratively motivated.

First, they obtain a list of words semantically associated with *A* and *B*, eventually using only the ones whose mutual association (*sōgō rensō*) score exceeds a given threshold. The authors differentiate between mutual association and similarity (*ruijido*), taking advantage of both categories. High level of similarity between *A* and *B* is used as a hint that their relation is rather literal than figurative. It is expected that — if they are connected by metaphorical mapping — *A* and

B should belong to different conceptual domains.

The authors use the semantic information provided by Nihongo Goi Taikei[5] — a large-scale thesaurus of Japanese language, as well as popular morphological analyser MeCab[6] equipped with NEologd dictionary[7].

Using a number of heuristics (e.g. if *A* in expression *A no B* is a formal noun *keishiki meishi*, treat the whole expression as literal), the authors perform a classification experiment. The model's performance is tested on 200 instances split evenly into literal and metaphorical classes. The results reported by the authors are: precision 78%, recall 75%, and F1-score 76%.

Hashimoto et al. [39] propose a lexical knowledge base for idioms recognition. The authors emphasize that AI's inability to identify idioms often results in mistranslation. They illustrate the issue with the Japanese sentence *Kare wa mondai o kaiketsu ni hone o otta* 'He struggled to solve the problem' erroneously translated into English as *He broke his bone to the resolution of a question*. In this case, machine translation algorithm failed to recognize *hone o oru* 'to make efforts; to take pains' (lit. 'to break a bone') as being used figuratively.

One of the causes of erroneous automatic translation indicated by the authors is the transformation of the fixed phrases. It can take shape of adding a word in between phraseme's constituents (e.g. *kare wa yaku ni tatsu* 'he is helpful' vs. *kare wa*

[5] http://www.kecl.ntt.co.jp/icl/lirg/resources/GoiTaikei/; last accessed on 1 March 2022.

[6] http://taku910.github.io/mecab/ last accessed on 1 March 2022.

[7] https://github.com/neologd/mecab-ipadic-neologd last accessed on 1 March 2022.

yaku ni sugoku tatsu 'he is very helpful'), conjugation of such constituents (e.g. *kare wa yaku ni tatta* 'he was helpful'), and so forth.

As pointed out by the authors, idioms often do not allow for morphological modifications (passivization, causativization, etc.). In consequence, any change within the inner structure of a multi-word entity may be a hint for its literal meaning. For example, in *kare ga hone o orareta* 'he got his bone broken [by someone/something]', the figurative reading is not possible. On the other hand, causative *kare ni hone o oraseru* can be understood in both literal 'make him break a bone' and figurative 'make him struggle/make effort' senses.

The authors divide idioms into three categories:

1. Non-transformable and non-ambiguous (e.g. *mizu mo shitataru* 'extremely handsome'),
2. Transformable and non-ambiguous (e.g. *yaku ni tatsu* 'to be helpful'),
3. Transformable and ambiguous (e.g. *hone o oru* 'to make efforts').

Tested on the data comprising idioms belonging to the third class, a rule-based classification algorithm presented by the authors achieved F1-score of 80%. The model sometimes failed to distinguish true senses of ambiguous expressions, whose interpretations depended on a larger context.

In [38], Hashimoto & Kawahara present the results yielded by the improved version of the algorithm described above. In

this trial, the authors utilize a large corpus of sentences collected from the Internet. They gather 146 ambiguous idioms (e.g. the above-mentioned *hone o oru* 'to struggle; to make effort', lit. 'to break a bone' or *goma o suru* 'to flatter', lit. 'to crash sesame'). The authors take advantage of both common Word Sense Disambiguation features (e.g. hypernyms of the context words, etc.) and idiom-specific features (those related to morphological constraints). The proposed algorithm achieved accuracy scores of 89.25% and 88.86%, depending on the features used.

The methods presented in this subsection aimed at identifying similes, metaphors with specific syntax, and idioms, respectively. None of them was designed to search for *all* types of figurative language. This motivated the approach I introduce in Chapter 3.

2.3.2 English as the Target Language

First approaches to automatic metaphor identification in English date back to the early 1990s [25]. A review of historically important works in the field can be found in [92, pp. 76-82].

As shown in the related survey [88], over the last few years much popularity was gained by metaphor detection using deep learning. To some degree, this growth in interest can be attributed to the emergence of language models like BERT (Bidirectional Encoder Representations from Transformers) [21] achieving state-of-the-art performance irrespective of the

NLP task they are being used for. This tendency can be noticed by looking at the list of models participating in the second Metaphor Detection Shared Task [60], where almost all of them use the implementations of ELMo (Embeddings from Language Model) [83], BERT, or some of its derivatives, such as RoBERTa (Robustly Optimized BERT Pretraining Approach) [62] or ALBERT (A Lite BERT) [58].

Presumably, this constatation led Neidlein et al. to publish their analysis [73] of recent metaphor recognition systems based on the language models. The authors argue that although the new models yield very satisfactory results, their design often shows considerable gaps from a linguistic perspective, indicated by the fact that they perform substantially worse on unconventional metaphors than on conventional ones. Subsequently, they present another finding that should be of great value to the whole community. First, the reader should know that VUAMC (Vrjie Universiteit Amsterdam Metaphor Corpus) [100] is the corpus underlying the two most frequently used datasets in metaphor detection research, specifically VUA-ALL-POS (All Parts of Speech) and VUA-SEQ (Sequential). Neidlein et al. reveal in their paper [73] that in recent research, some authors compare their results achieved using VUA-SEQ to the results gained on VUA-ALL-POS. As they point out, the underlying corpus remains the same, but the VUA-SEQ is substantially easier to do well on than VUA-ALL-POS, and thus such comparisons are inherently unfair. Later in the paper, the authors present results

for both VUA-SEQ and VUA-ALL-POS using a number of models published by other researchers and their own model as well. Implementation of their method achieves an F1-score of 77.5% on VUA-SEQ and only 69.7% on VUA-ALL-POS, which indeed proves that conflating these two cannot be considered a good practice.

DeepMet [104] is the winner of the second Metaphor Detection Shared Task mentioned before. It managed to outperform all the other models on any given data subset, often by a large margin. DeepMet uses RoBERTa as its structural foundation, and siamese architecture with two Transformer [107] encoder layers to process different features. The authors reformulate a metaphor detection task from a classification or sequence labeling problem to a reading comprehension task. DeepMet utilizes five categories of input features: global text context, local text context, query word, general POS (parts of speech), and fine-grained POS generated using SpaCy[8]. The overall performance is boosted by using ensemble learning and metaphor preference parameter α, helping the model to achieve a better recall score. This parameter is introduced due to the fact that the metaphor datasets are highly unbalanced, meaning they comprise many more target words belonging to the non-metaphorical class.

Similar to my approach, in the work of Wan et al. [109],

[8] https://spacy.io/; last accessed on 1 March 2022.

dictionary definitions are also used to improve the performance of the proposed BERT-based model. It is noteworthy that the authors perform not only metaphor detection but metaphor interpretation as well. In order to do so, they utilize every definition of the given word that is available in the Merriam Webster dictionary[9]. Using attention, they try to select the one that is semantically closest to the target word's contextual meaning. Afterwards, they concatenate all of the definitions' representations with the contextualized representation of the target word. Authors test their method on 3 datasets: VUA-SEQ, TroFi (Trope Finder) [8], and PSU CMC (PSU Chinese Metaphor Corpus) [64]. They use BERT for the experiments on TroFi and VUA-SEQ, and the Chinese BERT for PSU CMC.

Although both my and Wan et al.'s models use lexical information as one of the input features, there are several dissimilarities that make our approaches fundamentally different. First, while I collect all of the definitions in a simple and completely automatic manner, Wan et al. recruit a number of annotators to improve a part of their datasets. Even though both myself and the authors use dictionaries as the source of the word descriptions, the reason I do it and the goals I am trying to achieve are very different. While I am trying to extract the target word's meaning in its most basic literal sense regardless of the surrounding context, Wan et

[9] https://www.merriam-webster.com/dictionary/; last accessed on 1 March 2022.

al. search specifically for the definition semantically closest to the contextual meaning. Another dissimilarity lays in the choice of the dictionary. While Wan et al. choose Merriam-Webster, I am using the openly available Wiktionary[10], which also grants me with a larger number of definitions to utilize. As of November 2021, Wikitionary included 809,475 gloss definitions and 1,300,634 definitions in total just for English[11].

I could not find the precise information on how many definitions there are in the version of Merriam-Webster used by Wan et al., but the explanation found on the dictionary's website could hint at being somewhere over 275,000[12].

The work that has become the main inspiration for my project is that from Choi et al. [16]. The authors present MelBERT (Metaphor-aware late interaction over BERT), the model for metaphor detection using RoBERTa as its architectural foundation. MelBERT's design allows for using the principles of MIP (Metaphor Identification Procedure) simultaneously with the concept of SPV (Selectional Preference Violation), both of which I describe in detail in Section 4.2. MIP was proposed by the Pragglejaz Group [36] and provides the reader with instructions on how to establish whether a word is being used figuratively or not in a given context.

[10] https://en.wiktionary.org/wiki/; last accessed on 1 March 2022.

[11] The latter includes the number of definitions for *inflections, variants, and alternative spellings* as well, cf. https://en.wiktionary.org/wiki/Wiktionary:Statistics; accessed on 21 December 2021.

[12] This is the approximate number of word choices available in Merriam-Webster's Collegiate Thesaurus API; Merriam-Webster's Collegiate Dictionary API features more than 225,000 definitions, cf. https://dictionaryapi.com/products/index; last accessed on 1 March 2022.

The concept of Selectional Preference Violation in relation to metaphor was brought to the attention of computational linguistics by Wilks [26] and it can be said to focus on the degree of semantic *compatibility* between senses of given lexical units. Utilizing both strategies together, while using additional features such as POS tags, as well as local and global context, Choi et al. conduct a set of experiments on multiple datasets recognized in the field of metaphor detection: MOH-X (Mohammad et al. [2016] dataset) [72], TroFi, and two datasets based on VUAMC (for details on the datasets cf. Section 4.4). Subsequently, their results are compared to those achieved by some of the strongest benchmarks, including Su et al.'s DeepMet [104], which was briefly introduced above. While MelBERT is not the first model following the guidelines of MIP or SPV in metaphor detection, I found the fact of complementing linguistic theory with the power of recently published bidirectional[13] language models appealing and decided to further build on MelBERT's authors' ideas. This sort of holistic approach seems to also address the issue of having too much focus on technical innovations while disregarding related linguistic theories by the authors of recent models designed for metaphor identification, as signalized by Neidlein et al. [73]

[13]Or *non-directional*, as this would arguably be a more appropriate description, cf. [15].

Chapter 3

Metaphor Detection Experiments in Japanese

In this chapter I present the results of two classification experiments conducted using Japanese as the target language.

While in both experiments I use the same datasets, in each of them I extract different kinds of input features. In the first experiment, I take advantage of the seed words whose choice was inspired by the cognitive-linguistic perspective on metaphor. In the second one, I use the bag-of-words technique to represent the input. In both experimental trials I compare the performance of a number of standard machine-learning-based classification algorithms available via scikit-learn framework [82].

Unlike in the trials conducted on English data (see Chapter 4), in the experiments presented in this chapter I use a whole sentence — not a single token — as the unit for classification. This decision is motivated by the fact that setting a border between parts of sentence used literally and non-literally is often unfeasible. There are cases where such demarcation line is unambiguous, as in:

– *The land belongs to the crown,*

where only *crown* appears to be used in a non-literal sense. It is a conventional and readily recognizable example of metonymy. However, similar lack of ambiguity is not always the case, which can be portrayed with the following sentences:

– *Light is but the shadow of God.* [9, p. 274]

– *His theory has thousands of little rooms and long, winding corridors.* [56, p. 53]

– *Bō hakushi no tateru riron wa doremo genkanguchi ga kazari-taterarete ite, haitte iku no ni jikan ga kakaru.* [113, p. 57] 'The entrance gates to all of the Doctor X's theories are full of decorations and it takes time to get inside.'[1]

I argue that, in similar sentences, trying to single out all of the individual words used figuratively would be extremely difficult, even for the human annotator. It follows that an attempt to force the algorithm to carry out such an operation would be counter-intuitive.

3.1 Task Description

In the two classification experiments described in this chapter, the task was to automatically identify sentences comprising figurative expressions (see also pp. 19–20 in Section 1.4), using different machine learning algorithms. In both experiments, vector representations of 82 test sentences were passed to the classifiers, whose objective was to correctly predict the label

[1] English translation by myself.

(non-literal or literal) of a given input vector. Two kinds of input features used in the experiments are described further in Sections 3.4.1 and 3.5.1.

3.2 Datasets

Training data prepared for the classification experiments comprised 25,947 sentences used figuratively and the corresponding number of sentences with supposedly literal senses. *Supposedly*, because — while no genre is completely free of figurative language — the sentences excerpted from encyclopedic descriptions, news articles, and transcripts of formal discussions seemed likely to be used in literal senses.

The first group was obtained from the digitalized dictionary of figurative expressions edited by Onai [76]. The second was collected using three different sources: Japanese Wikipedia dump [46], articles from Livedoor News [63] and local parliamentary minutes [45]. As for the dictionary of figurative expressions, my laboratory was provided with its abridged electronic version not available on the market, originally comprising 25,979 items.

Testing data was initially composed of a modest number of 50 sentences randomly taken from the texts of Japanese novels openly available at Aozora Bunko digital library [2]. In the process of annotation described in greater detail in the following section, 41 sentences were labeled as literal and only 9 as non-literal. In order to make these numbers equal,

another 32 sentences with figurative senses were transferred from the training set to the evaluation set.

Eventually, 51,894 sentences (25,947 figurative and 25,947 literal) were utilized for training and 82 sentences (41 figurative and 41 literal) for testing.

3.3 Annotation

Three Japanese native speakers were asked to annotate 50 sentences obtained randomly from Aozora Bunko digital library [2]. They were instructed to decide whether a sentence includes any figurative expression, regardless of its type. Instruction was prepared in Japanese, and it can be translated into English as follows:

> Read the following sets of three sentences one by one. Try to decide whether any figurative expression (metaphor, simile, etc.) is used in the middle (the second) sentence and choose the correct answer from the parenthesis. If you think that there is a figurative expression in use, indicate this word or phrase using an underline.

As shown by Dobrzyńska [23], the evaluation of whether certain expression is used figuratively or literally is often context-dependent. I have therefore provided the annotators with the adjacent sentences as well (the previous and the next to the target one). Employing an odd number of annotators prevented the anticipated agreement-related problems: sentences were labeled according to the majority's decision.

Here, I briefly explain some of the difficulties accompanying the annotation process. The expression *doko made mo* (lit.

'anywhere', 'wherever') from the sentence: *kenkyū wa doko made mo kenkyū de aru* 'a research is a research, from A to Z' has been labeled as literal by all three annotators. To my understanding, this expression can be viewed as based on the underlying spatial metaphor.

Another expression, in my opinion, mistakenly labeled as used in a literal sense is the following: *jikaku ga tsuyoku natta* '[one] became more aware' (lit. 'awareness became strong'). This is an example of personification, since *awareness* is not a living entity and therefore cannot become *strong* in a literal sense.

Similarly, the sentence *Haisha wa* (...) *yagate shiorete suwarimashita* 'After a while, the stomatologist became downcast and took a sit' (lit. 'After a while, the stomatologist withered and sat'), while apparently including a metaphorical expression (for people do not wither as plants), was labeled as literal by one of the annotators.

In another sentence, although figurative use of language is highlighted by the presence of a simile marker *yōna*, the expression *mune o utareru yōna kōkei* 'a touching scene' (lit. 'sight as if chest was hit [by something]') was classified as literal by one of the annotators.

Longer sentences sometimes included multiple linguistic units used figuratively. In the following example, several expressions conveying non-literal senses can be found: *Hitode wo karizu, fūfu dake de mise o kirimawashita node, yoru no jūji kara jūniji koro made no ichiban tatekomu jikan wa me*

no mawaru hodo isogashiku, shōben ni tatsu hima mo nakatta
'Because me and my wife were running the store (lit. 'cutting the store around') without anyone's help (lit. 'not borrowing a hand'), at the rush hour from 10 o'clock to around 12 o'clock in the evening it was so busy, that my head was spinning (lit. 'busy to the extent eyes were spinning') and I didn't even have time to take a leak'.

3.4 Experiment 1

3.4.1 Features

It appears as hardly possible to coin the definition of figurative language, that would be free of contradictions and sufficient enough to cover all of its possible manifestations. It seems then reasonable to search for the non-literal use of language, enumerating at least some of the features conceivable as non-trivial in its identification. Inspired mostly by the research on metaphor within the field of cognitive science, I propose seven groups of seed words, whose presence in the sentence may hint at its non-literal usage.

First category comprises the words denoting body parts (BP), as they are frequently encountered in figurative expressions [65]. In the experiment, I use 57 of such words. Their sample can be viewed in Appendix 1 provided at the end of this dissertation.

Due to their specific syntax, *similes* (SI) are relatively easy to identify in text. As they belong to the broad category of

figurative language, I decided to treat the words indicating presence of similes as another input feature. Their list consists of 24 elements; see Appendix 2 for a sample.

Orientational metaphors GOOD IS UP and BAD IS DOWN are among the most well-known conceptual metaphors discovered. Expressing moral judgements, we often speak in terms primarily denoting directions and positions in space [115]. I use the words indicating *spatial positions* (PI) as another feature for the model. Their list comprises 17 elements; see Appendix 3 for a sample.

Abstract concepts are often described metaphorically with the words related to sensory impressions [69]. For this reason, I use the names of *colors* (CO) as the next feature. I supply the list of the traditional Japanese colors with a small number of modern words belonging to the xeno-Japanese sublexicon *gairaigo* (*pinku* 'pink', *orenji iro* 'orange' and *gurē* 'gray'). In this group I have gathered 242 elements, whose sample can be found in Appendix 4.

It is well known that metaphor is often employed as a means to describe feelings [24, 78]. Conceptual metaphors such as LOVE IS A JOURNEY and ANGER IS A HOT FLUID IN A CONTAINER can serve as the examples. I have constructed a list of 266 words related to *emotions* (EM). Their sample can be seen in Appendix 5.

I have also prepared a list of 130 most frequent words found in the dictionary of figurative expressions [76]. I name them *frequently poetical* (FP) and use as another feature. Their

sample can be viewed in Appendix 6.

As Uchiyama & Ishizaki [106] point out, it is often the case that the Japanese compound verbs of shape V1-V2 (e.g. *yomi-naosu* 'reread', *kaki-ageru* 'finish writing/drawing', etc.) are used metaphorically. I have constructed a list of verbs belonging to the lexical subsystem *wago* and use them as the last feature (WV). The list consists of 1,044 elements; see Appendix 7 for a sample[2].

Each of the 51,976 (51,894 + 82) sentences available in the datasets is transformed into an array composed of seven elements corresponding to the seven categories of features introduced above. The algorithm checks, whether any member of the given feature group is present in the given input string. If at least one of the seed words belonging to a specific list is found, the value of 1 is assigned to the respective place in the input feature vector. If no such word is found, the value of 0 is assigned.

As mentioned, seven types of features (*body parts*, *simile indicators*, *spatial positions*, *colors*, *emotions*, *frequently poetical & wago verbs*) have been adopted as possible cues for figurative use of language. In order to compare the effectiveness of every arrangeable combination of the features, the program is run 127 times.

As there are only two types of labels, the algorithm's objective is to output a value of 1 for the input representing

[2]The complete versions of the appendices are available at: http://arakilab.media.eng.hokudai.ac.jp/MetaAppendix.zip; last updated on 1 March 2022.

non-literal sentence, and a value of 0 otherwise. A given prediction is counted as correct if the predicted value matches the one determined by the annotators.

The experiment is conducted in Python, using scikit-learn framework [82]. The performance of the following classification algorithms is measured and compared: Random Forest, Naïve Bayes, Logistic Regression, K-Nearest Neighbors, Linear Discriminant Analysis and Support Vector Machines. The values of the hyperparameters are set to default with the exception for SVM's kernel type (sigmoid kernel is selected).

3.4.2 Results

Figures below (3.1, 3.2, 3.3, 3.4, 3.5, 3.6) allow for the comparison of the results (expressed in F1-score) achieved by each of the selected classification algorithms using different subsets of features.

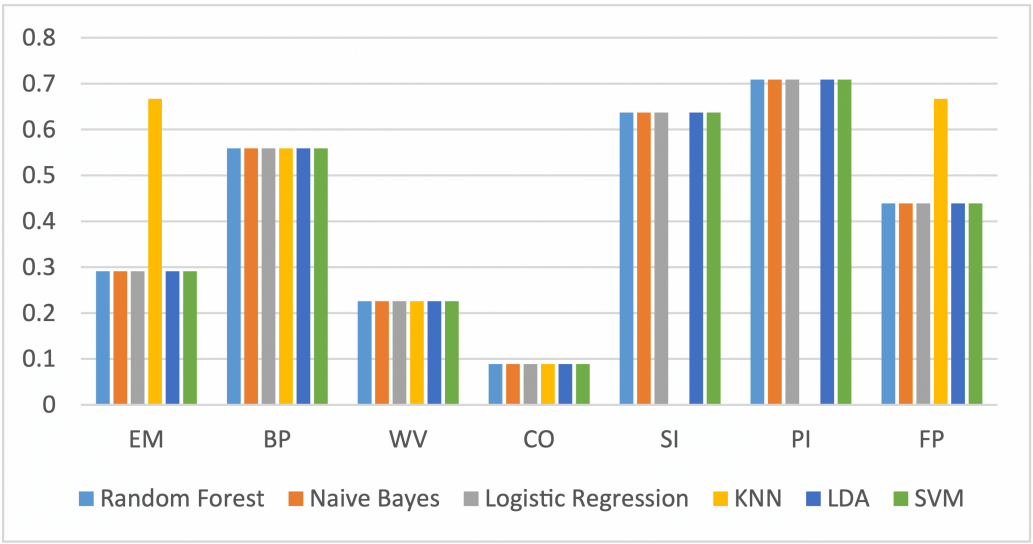

FIGURE 3.1: Results (F1-score) for a single feature

Figure 3.1 shows that — when compared in isolation — out of all the feature groups, it is the *position indicators* (PI), which allows five of the six tested classifiers to achieve the highest F1-score (70,9%). Using the seed words belonging to the categories of *emotions* (EM) and *frequently poetical* (FP), K-Nearest Neighbors outperforms other algorithms, reaching level of 66,7%. In this setup, other algorithms receive scores of 29,1% and 43,9% for EM and FP, respectively.

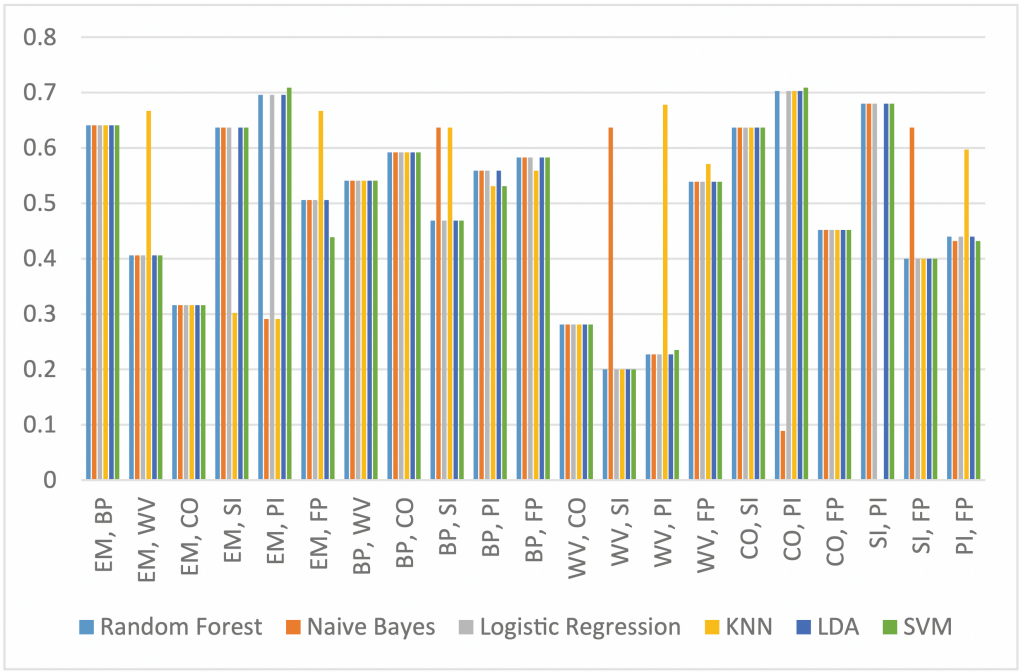

FIGURE 3.2: Results (F1-score) for two features

The trials conducted using different subsets of two features (cf. Figure 3.2) confirm that KNN often predicts differently than other algorithms and works visibly better in certain setups (WV & PI, PI & FP, BP & SI, EM & PI, EM & WV). However, the highest F1-score of 70.9% is yielded by SVM using two combinations of features (EM & PI and CO & PI).

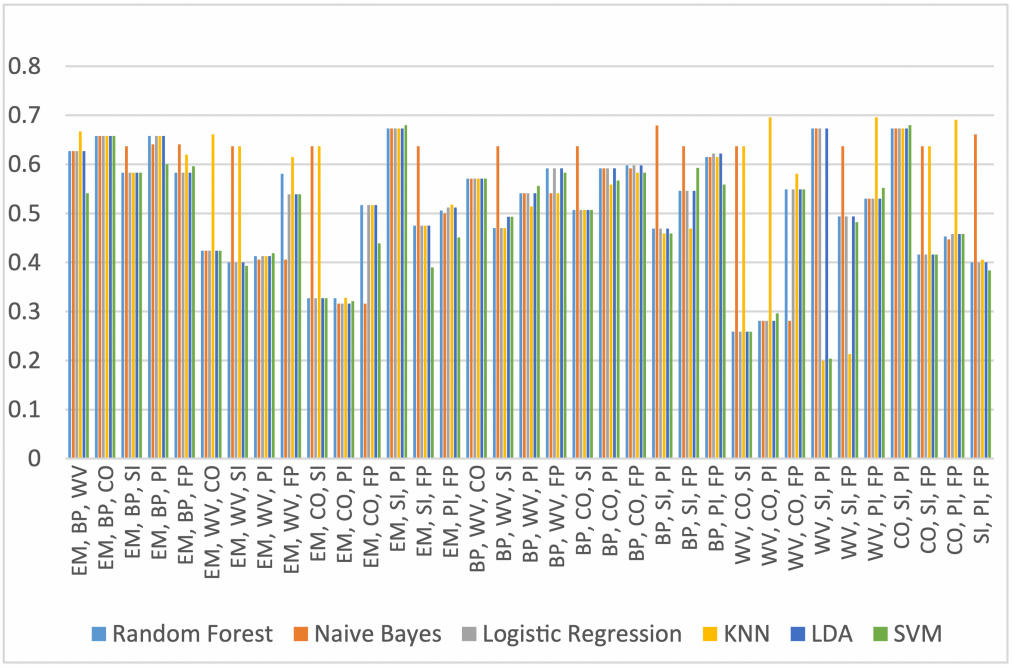

FIGURE 3.3: Results (F1-score) for three features

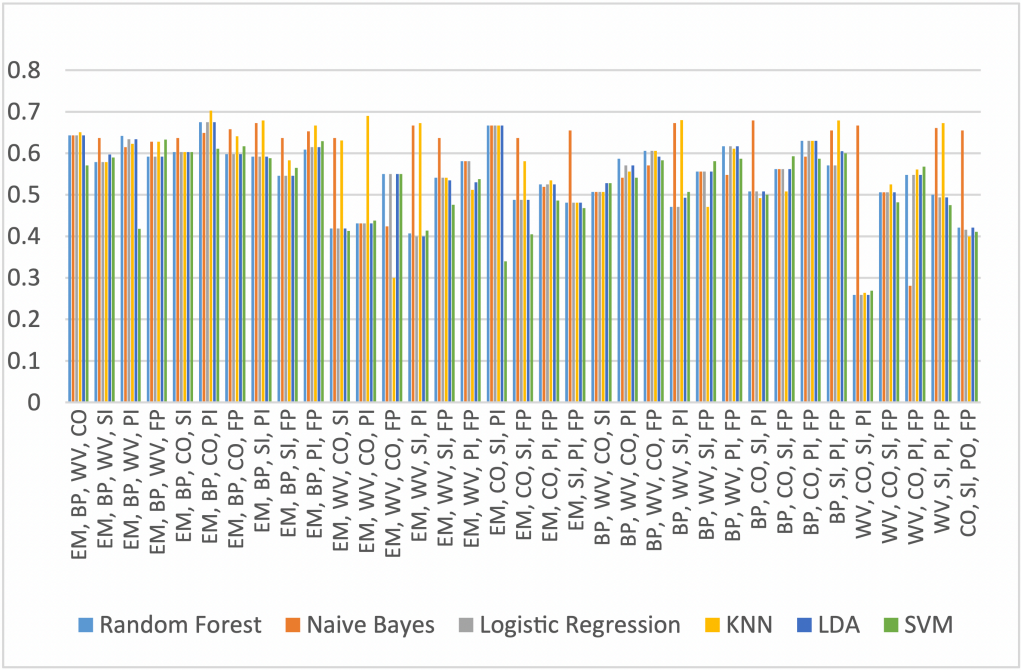

FIGURE 3.4: Results (F1-score) for four features

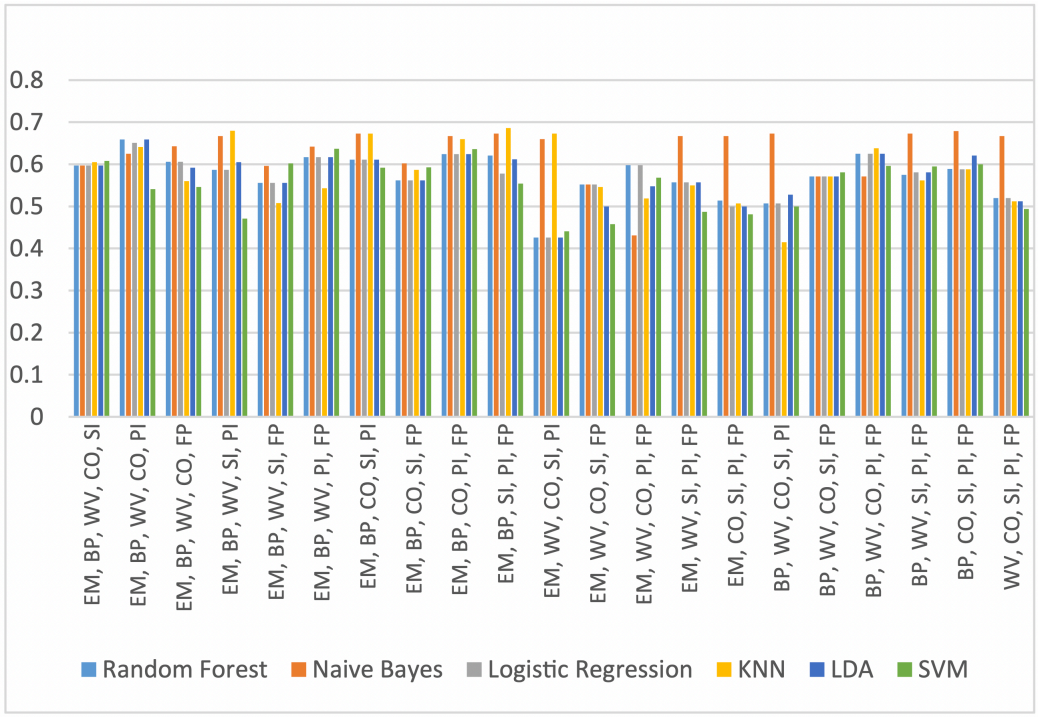

FIGURE 3.5: Results (F1-score) for five features

Figures 3.4 & 3.5 once again demonstrate that KNN tends to make different decisions than most classifiers, which allows it to achieve better scores. Trials with the subsets of four and five features show that, in such settings, Naïve Bayes is often superior to other algorithms.

The results of trials using six features (Figure 3.6) show some decrease in scores of KNN and Naïve Bayes. Using the subsets of six features makes the margin separating the score achieved by different algorithms less significant.

Figure 3.7 presents the results of the trials with all of the seven kind of features utilized. In this setting, Support Vector Machines achieves the best F1-score (64,4%). The scores yielded by generally well-performing KNN become the lowest relatively to other algorithms.

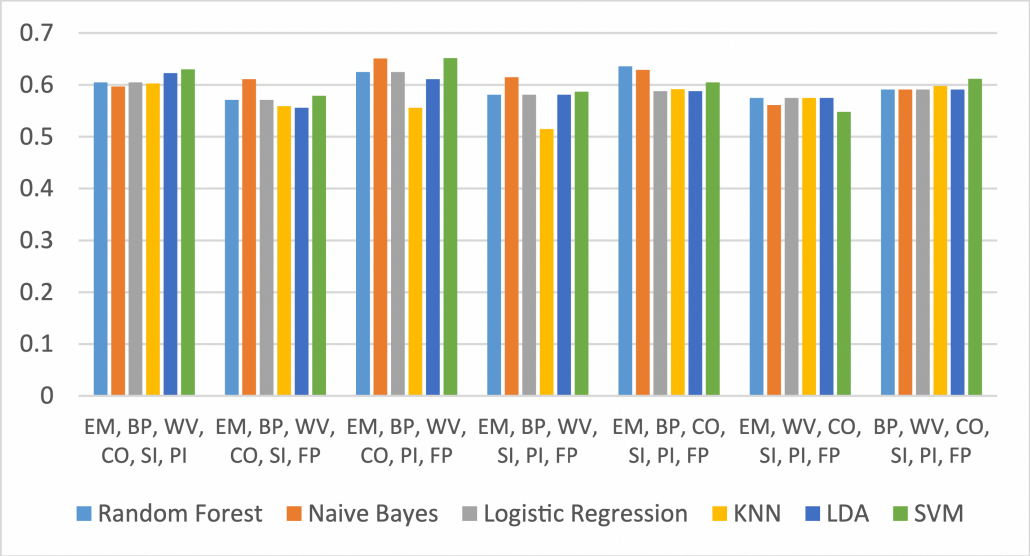

FIGURE 3.6: Results (F1-score) for six features

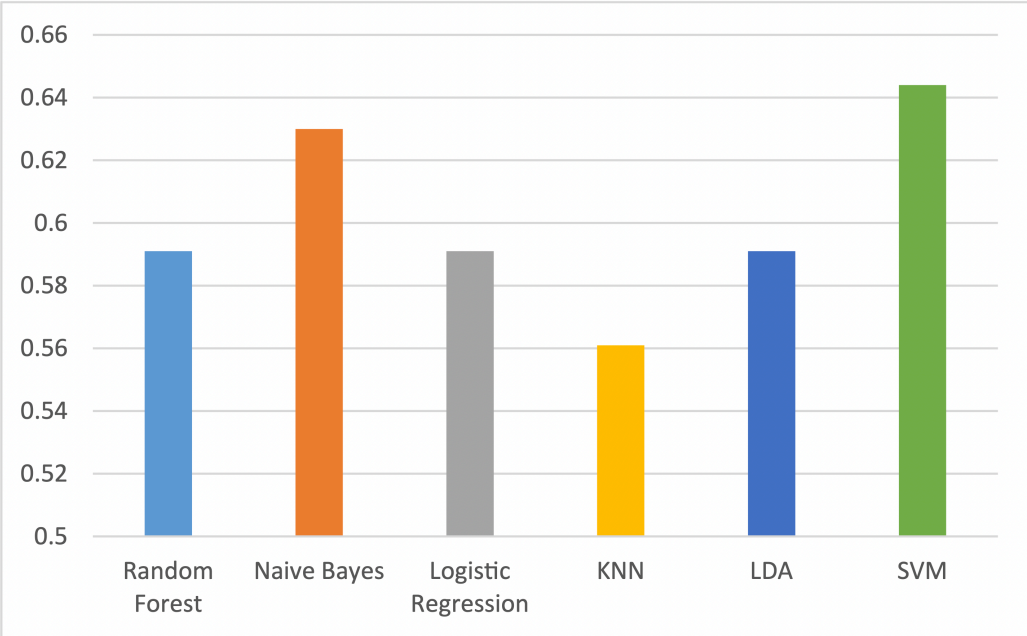

FIGURE 3.7: Results (F1-score) for all seven features

3.4.3 Considerations

Some adjustments in the current shape of the proposed features seem desirable.

For example, more rigorous approach to *wago verbs* (WV) may be relevant. Specifically, selecting a smaller number of seed words statistically more likely to occur within metaphorical expressions may improve the final results.

Distinguishing between literal comparisons and similes could elevate the algorithm's performance when using the *simile indicators* (SI) feature. Whether the compared elements are in hyponymy relation — which may suggest a literal meaning of the expression — can be established by using knowledge bases like WordNet [28].

As the concepts of animals are frequently employed by figurative expressions [79, 96] (e.g. the proverb *ken'en no naka* 'like cat and dog', lit. 'dog-and-monkey relationship'), it may be beneficial to use the words denoting members of this semantic field as another category of features.

Error analysis has shown that the sentences comprising *spatial position indicators* (PI) tended to be predicted as literal. It might be due to the type of pattern matching adopted for this experiment. The ideograms primarily denoting directions and spatial loci can be found in many compound words, including those usually not recognized as figurative. For example, the ideogram 中, read as *chū*, *jū* or *naka*, and used primarily in the sense of 'middle; center; medium', appears in

the compound words: *Chūgoku* 'China', *chūko* 'used', *senaka* 'back of the body', which — even if historically motivated by metaphors — nowadays are not seen as figurative units.

This kind of generous approach was taken in order to allow for detecting conceptual metaphors underlying a single word's structure. For example, *gehin* 'vulgar; indecent' and *buka* 'a subordinate; a henchman' comprise the ideogram 下 (here read as *ge* and *ka*, respectively) whose primary sense 'down; bottom; below' indicates that the words in question are motivated by the conceptual metaphor BAD IS DOWN. The already mentioned 中 *chū* can be suffixed to other words (e.g. *denwa* 'telephone', *kaigi* 'meeting', *shigoto* 'work'), conveying a metaphorically motivated sense 'to be in the middle [of some activity]'.

3.5 Experiment 2

3.5.1 Features

In the second experiment, the word-count vectors of 32,231 dimensions are constructed for each of 51,976 input sentences (51,894 and 82 items from training and testing sets, respectively). The number of dimensions is equal to the number of unique words available in the training set after removing simile indicators and stopwords listed in Table 3.1. Incorporating to the dictionary even the least frequent members of the vocabulary available in the training set is motivated by the

TABLE 3.1: Simile indicators and stopwords removed during preprocessing.

Simile Indicators	**Stopwords**
yō, *yōna*, *yōni*, *yōda*, *fū*, *fūna*, *fūni*, *fūda*, *mitai*, *mitaina*, *mitaini*, *mitaida*, *gotoki*, *gotoku*, *gotoshi*, *marude*, *ppoi*, *kurai*, *gurai*, *hodo*, *sazo*, *samo*, *sanagara*, *ikanimo*, *atakamo*	*suru*, *itasu*, *iru*, *oru*, *aru*, *naru*, *kore*, *sore*, *are*, *kono*, *sono*, *ano*, *kara*, *wakeda*, *koto*, *mono*, *nado*, *nani*, *keredomo*, *demo*, *watashi*, *da*, *noda*, *nda*, *nado*, *tai*, *dai*, *na*, *ni*, *nu*, *ne*, *no*, *tsu*, *to*, *ya*, *ga*, *wa*, *o*, *go*, *de*, *mo*, *ka*, *yo*, *sa*, *ru*, *e*, *reru*, *rareru*, *seru*, *masu*, *desu*, *gozaimasu*

assumption that figurative use of language may be indicated by the occurrence of a rare word within the input sentence.

As a preliminary step towards conducting the experiment, all sentences from both training and testing sets are parsed with JUMAN [47]. As this format is not used for encoding standardized Japanese text, all ASCII characters are discarded. Punctuation marks are removed as well.

Most of the sentences available in the dictionary of figurative expressions [76] include similes. Specifically, at least one of the simile indicators listed in Table 3.1 appears in 24,641 of 25,947 metaphorical sentences comprised by the training set. In contrast, this is the case in only 1,616 sentences labeled as literal. To prevent inducing a prediction bias, all of the words listed as simile indicators in 3.1 were removed from both non-literal and literal sentences used for training.

Three standard classification algorithms, namely Naïve Bayes (multinomial), Support Vector Machines (Linear SVC), and Random Forest Classifier are used via scikit-learn library [82]

for Python. The type of SVM's optimization problem is set to *primal* and its maximum number of iterations is set to 2,000; all other hyperparameters are used with their default values.

3.5.2 Results

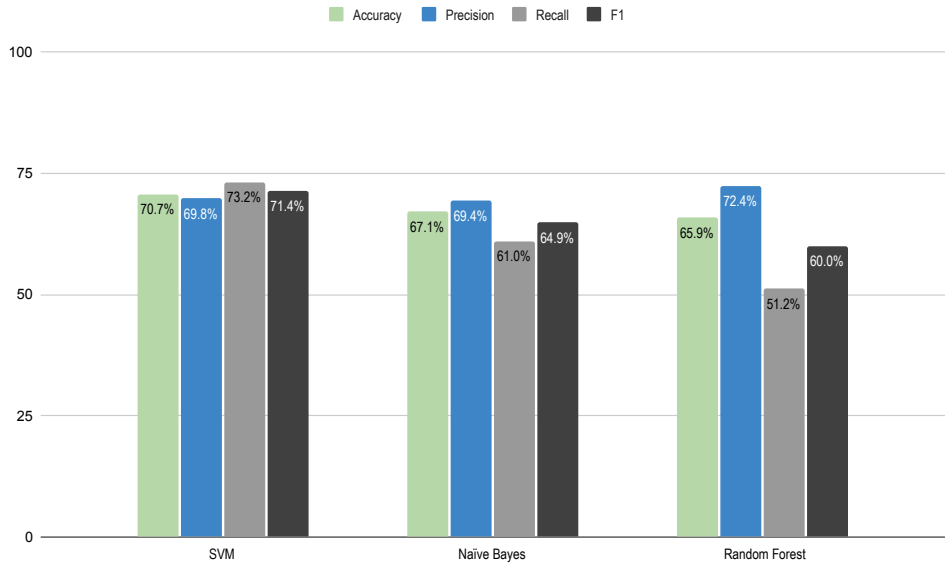

FIGURE 3.8: Bag-of-Words: Classification Results

As can be seen in Fig. 3.8[3], SVM achieves the most balanced scores with precision of 69.8%, recall of 73.2%, and F1-score of 71.4%. Overall the worst results are yielded by Random Forest. Although, in terms of precision, it outperforms the other two algorithms with the score of 72.4%, otherwise it falls behind with recall of 51.2%, and F1-score of

[3]The results presented in this section are different from the original ones shown in the paper released in 2019 [3]. At the time of the paper's publication, in the original code, negative and positive labels were reversed (i.e. metaphorical sentences were labeled with zeros and literal sentences with ones), which has been fixed. I have also applied minor changes to the lists of stopwords and simile indicators, which are displayed in Table 3.1, and modified the pattern matching method. The currently presented results are based on the code including said modifications.

60.0%. Naïve Bayes performs overall better with precision of 69.4%, recall of 61.0%, and F1-score of 64.9%.

3.5.3 Considerations

The results presented above suggest that even relatively simple method based on the bag-of-words model can be useful in figurative language detection. The F1-score (71.4%) of the best-performing Support Vector Machines not only exceeds the highest of the corresponding values yielded in the first experiment (70.9%), but the achieved results are overall much more balanced as well. Specifically, the highest F1-score yielded in the first experiment was derived from precision of 56.5% and recall of 95.1% (accuracy was 61%). In the current trial, all of the scores yielded by SVM reach similar values close to 70%, which shows greater stability displayed by the algorithm.

As stated, the results show that even rather uncomplicated term-frequency-based method can grant a certain level of utility in the analyzed task. Nonetheless — being order-agnostic and context-unaware — this method displays some undeniable technical limitations as well. As already mentioned in Section 3.3, the evaluation of certain expression's metaphoricity can largely depend on its context (for example, consider the possible interpretations of *all men are animals*). To improve the performance, it may be appropriate to adopt a method allowing for utilizing the information provided by the context.

There was a number of obstacles I have encountered in my approaches to automatic metaphor identification in Japanese described in this chapter.

One of them was a side-effect of collecting the training data in an unsupervised manner. Specifically, obtaining a large number of sentences *entirely* and unarguably literal proved unfeasible. The analysis of the automatically retrieved training data showed that even encyclopedic descriptions are not free from figurative language.

Another major issue is related to the annotation performed on the testing data. As mentioned in Section 3.3, the annotators were not recognizing the figurative nature of certain expressions, especially conventional metaphors highly frequent in everyday language. Similar problem was encountered in one of my earliest approaches to metaphor processing in Japanese [6]. It exposed a high level of disagreement between the two employed annotators, where 40% of times their evaluation was different (κ coefficient = 0.297). Among the expressions labeled by one of them as literal were the following: *jijō to musubitsuita* 'circumstances-related' (lit. 'tied to circumstances'); *ki o momanakute wa naranai* 'cannot help but to worry' (lit. 'cannot help but to rub one's ki [one's internal energy]'), and *uwasa wa uwasa o umu* 'rumor feeds upon rumor' (lit. 'rumor gives birth to rumor'). In order to bypass the disagreement issue, in this approach I decided to employ an odd number of annotators and follow the majority's decision. This, however, did not prevent the expressions

like *jikaku ga tsuyoku natta* '[one] became more aware' (lit. 'awareness became strong') from receiving non-metaphorical labels.

Perhaps the instructions the annotators were provided with were not detailed enough. In consequence, there might have been too much space left for the intuition-based interpretation. In [13], Caballero & Antuñano argue that:

> metaphoricity may be seen as a matter of degree: not all metaphorical language is regarded as such by all people, underlining the role of context and social convention in metaphor awareness and identification.

While this is true, a comprehensive set of annotation guidelines should allow for a more reliable output. This has been proven by the success of Metaphor Identification Procedure (MIP) [36], whose detailed description is provided in Section 4.2.1. Following the rules of its upgraded version — MIPVU [100] — Steen et al. compiled a large-scale VUA Metaphor Corpus (VUAMC)[4], which has become the most popular source of labeled data used for token-level metaphor identification in English.

Although — having modified them to address a number of cross-linguistic distinctions — it seems possible to utilize the guidelines provided by MIP in annotating Japanese datasets

[4]Details on the corpus annotation can be found at: http://www.vismet.org/metcor/documentation/annotation_procedure.html; last accessed on 1 March 2022.

as well, such an endeavor would largely exceed the scope of the current research[5].

To the best of my knowledge, in the field of Japanese NLP there is no commonly chosen dataset designed for metaphor identification that would allow for an objective comparison of different proposed methods in terms of their performance[6]. In contrast, already two editions of Metaphor Detection Shared Task utilizing mentioned VUAMC have been organized [61, 60].

In the following chapter I describe the details of my recently presented method for token-level metaphor identification in English. I use a number of openly available datasets, whose major part comes from VUAMC. This allows me to compare the performance of my newest model with several state-of-the-art methods proposed by other researchers.

[5]According to the information found in [100, p. 7], "the set of instructions was developed and tested over five years".

[6]From this perspective, the recently published Corpus of Japanese Figurative Language [50] seems to have the potential to become such a resource.

Chapter 4

Metaphor Detection Experiments in English

Recent years have brought an unprecedented and rapid development in the field of natural language processing. To a large degree this is due to the emergence of modern language models like BERT (Bidirectional Encoder Representations from Transformers) [21], XLNet [114], and GPT-3 (Generative Pre-trained Transformer 3) [12], which are pre-trained on a large amount of unlabeled data. These powerful models can be further used in the tasks that have traditionally been suffering from a lack of material that could be used for training. Metaphor identification task, which is aimed at automatic recognition of figurative language, is one of such tasks. The metaphorical use of words can be detected by comparing their contextual and basic meanings. In this chapter, I deliver the evidence that fully automatically collected dictionary definitions can be used as the optimal medium for retrieving the non-figurative word senses, which consequently may help improve the performance of the algorithms

used in metaphor detection task. As the source of the lexical information, I use the openly available Wiktionary[1]. My method can be applied without changes to any other dataset designed for token-level metaphor detection in English, given it is binary labeled. In the set of experiments, my proposed method — MIss RoBERTa WiLDe (Metaphor Identification Using Masked Language Model with Wiktionary Lexical Definitions) — outperforms or performs similarly well as the competing models on several datasets commonly chosen in the research on metaphor processing.

4.1 Task Description

A word-level metaphor detection task can be defined as a supervised binary classification problem, where a computational model predicts if the target word comprised by the input sentence is used metaphorically or not. The model is trained on datasets where each sample is composed of a number of features X and label y. Input features include the target-word, the sentence, and — in this method's case — the definition of the target word. The label is determined by the human annotator and takes values of either 0 or 1 (non-metaphor or metaphor, respectively). The model's goal is to learn the correlations between features and labels in the training set in order to correctly predict whether the given target word is

[1] https://en.wiktionary.org/wiki/; last accessed on 1 March 2022.

used metaphorically or not in the unseen input sentence from the testing set.

4.2 Metaphor Detection Procedures

In this section I present an overview of the two linguistic procedures for metaphor identification, which provide theoretical justification for the algorithmic architecture I adopt in my models as described in Section 4.3. Both of them have been successfully used in metaphor detection tasks over the years, becoming a popular choice among researchers in the field (cf. [25, 67, 93, 75, 112, 37, 30, 66, 16]).

4.2.1 MIP

MIP (Metaphor Identification Procedure) was introduced by Pragglejaz Group [36] and was designed as a method for identifying words used figuratively in discourse. As pointed out by the authors, the primary difficulty with metaphor detection is that researchers often differ in their opinions on whether a given word is used figuratively or not because of a lack of objective and universal criteria that could be applied to the task. Before MIP was proposed in 2007, the need for undertaking this problem had been signalized by other authors as well. For example, Heywood et al. [43, p. 36] suggested that:

> The fuzzy boundary between literal and metaphorical language can only be properly tackled by being maximally explicit as to the criteria for classifying individual expressions in one way or another.

Similarly, in their work dedicated to the analysis of metaphor in a specialized corpus, Semino et al. [91, p. 1272] admitted that:

> It seems to us that we still lack explicit and rigorous procedures for its identification and analysis, especially when one looks at authentic conversational data rather than decontextualized sentences or made-up examples.

Such concerns have become a motivation for creating a few-step procedure for metaphor identification that would be precise enough not to allow for much idiosyncrasy in the related decision making. In MIP, these steps include reading the whole text to understand its meaning, establishing the contextual meaning of each lexical unit in the text, determining whether a given lexical unit has a more basic sense in other contexts, and, finally, deciding if the contextual meaning contrasts with the basic meaning and can be understood by comparison with it. Since the explanation regarding the way of determining a word's basic meaning is of great importance to the current task, and paraphrasing inevitably leads to some information loss, I allow myself to directly cite the authors on this issue [36, p. 4]:

> For each lexical unit, determine if it has a more basic contemporary meaning in other contexts than the one in the given context. For our purposes, basic meanings tend to be:
>
> – More concrete; what they evoke is easier to imagine, see, hear, feel, smell, and taste.
> – Related to bodily action.
> – More precise (as opposed to vague).

– Historically older.

Basic meanings are not necessarily the most frequent meanings of the lexical unit.

Although this list might seem self-explanatory at first sight, in reality, establishing whether given meaning is basic often comes as no easy task. For example, there are cases in which a polysemous word has multiple senses from which one is historically older and the other is more concrete. In such a case, it is not clear which of the meanings should be considered the basic one. Consider the example sentence provided by the authors [36, p. 4] (emphasis added):

> For years, Sonia Gandhi has struggled to convince Indians that she is *fit* to wear the mantle of the political dynasty into which she married, let alone to become premier.

Here *fit* appears to be this kind of a word. As the authors write in their explication [36, p. 8]:

> The adjective *fit* has a different meaning to do with being healthy and physically strong, as in *Running around after the children keeps me fit*. We note that the "suitability" meaning is historically older than the "healthy"meaning; the Shorter Oxford English Dictionary on Historical Principles (SOEDHP) gives the "suitability"meaning as from medieval English and used in Shakespeare, whereas the earliest record of the sport meaning is 1869. However, we decided that the "healthy" meaning can be considered as more basic (using the description of "basic" set out earlier) because it refers to what is directly physically experienced.

This description suggests that establishing the basic meaning of a word, and thus deciding whether it is being used figuratively or not, still involves a certain amount of subjectivity. This problem is addressed by Steen et al. in [99, p. 771], where the authors suggest that the term *metaphor* should be in fact thought of as short for "metaphorical to some language user" as opposed to absolutely metaphorical.

4.2.2 SPV

Utilizing the concept of SPV (Selectional Preference Violation) as a tool in automatic metaphor recognition has been postulated most notably by Wilks [26, 111, 112]. Wilks argued that metaphors could be detected in a procedural manner by determining whether semantic preferences of linguistic units present in the sentence are not violated. To cite the author, such preference violation "can be caused either by some 'total' mismatch of word-senses (…) or by some metaphorical relation" [26, p. 182]. Such metaphorical relation may be illustrated with the famous Wilksian example: *My car drinks gasoline*, where the verb *drink* can be said to exhibit preference for an animate agent and a patient belonging to the semantic field of LIQUIDS. As it denotes an non-animate object, the noun *car* breaks the verb's selectional preference. Shutova et al. contrast this example with: "My aunt always drinks her tea on the terrace" [94, p. 310] in which selectional preference violation does not occur.

Metaphorical expressions breaking selectional preferences can be easily found in every-day language. Consider the sentence: "This new PlayStation is a beast". Since *PlayStation* is referring to a video game console, *beast*, defined as 'any animal other than a human', violates the subject's preference for an object belonging to the semantic field of MACHINES, which hints at the possibility of a metaphor being in use. In my work, similar to Choi et al. [16], I employ a rather simplified interpretation of the SPV concept. Specifically, as it is likely that it breaks some selectional preferences on the way, whenever some unexpected, unusual word occurs in a given context, I assume that it might be used figuratively. This is often the case and can be proven by the example above; if there was a rule that words can be used only in their literal senses, *device*, *console*, *machine*, or alike would be used in place of *beast*. Some other examples portraying this rule would be: "She took his life", where *backpack*, *wallet*, *sandwich*, etc. would be expected in place of *life*; "I smell victory", where *tomato soup*, *cigarettes*, or *gasoline* would have preference over *victory*; "They have been living in a bubble" where *house*, *mansion*, etc. would be used in place of *bubble*.

4.3 Model Structure

In this section, I present the architecture of MIss RoBERTa WiLDe, a model for Metaphor Identification using the RoBERTa language model. At its core, MIss WiLDe utilizes MelBERT

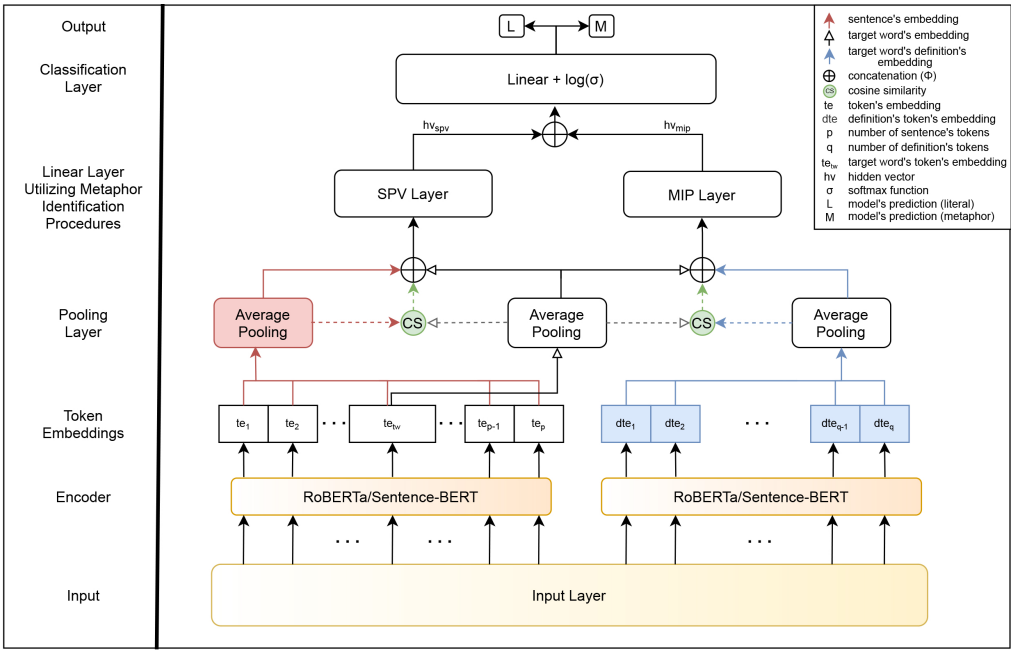

FIGURE 4.1: Model overview. Elements marked with colors other than black and white signify introduced innovations in relation to MelBERT [16]. Gradients stand for partial novelty.

(Metaphor-aware late interaction over BERT) published recently by Choi et al. [16] and therefore the architecture of the two models is almost identical. For the model overview, whose design was inspired by the aforementioned work, see Figure 4.1. Conceptually, MIss WiLDe and MelBERT take advantage of the same linguistic methods for metaphor detection, namely SPV (Selectional Preference Violation) and MIP (Metaphor Identification Procedure). While the implementation of the former in the proposed model remains mostly unchanged, the latter is affected by a different kind of input, which is the first novelty of my approach.

In order to determine whether the target word is used figuratively in the given context, Choi et al. utilize its isolated

counterpart as a part of the input. This is done for the same purpose MIss WiLDe takes advantage of the target word's definition (see Figure 4.2 and 4.3). Although using uncontextualized embedding of the target word proved to yield satisfactory results in the work of Choi et al., from a theoretical perspective, this approach does not seem to be free of flaws. Its main issue lies in the fact that during the pre-training stage, the word embedding is constructed by looking at the word's usages in various contexts. In consequence, at least some of these usages (even most of them, depending on the word in question) are already metaphorically motivated. On the other hand, using the target word's lexical definition — more specifically, using the first of the definitions listed in the dictionary — should be able to bypass this problem. This is because lexicographers tend to place the definition representing what is called the word's *basic meaning* at the top of the definitions list. Given the definition is indeed representing the word's basic sense, its embedding representation can be subsequently used to compare it with the contextualized embedding of the target word. If the gap between them is big enough, it can be estimated that the word is used figuratively.

Another motivation for using the definition rather than the target word itself is that — considering RoBERTa was pretrained on a large amount of text data with a sentence as its input unit — I anticipated that using sentences instead of single words might lead to some performance gains.

The second novelty of my method lies in the way in which

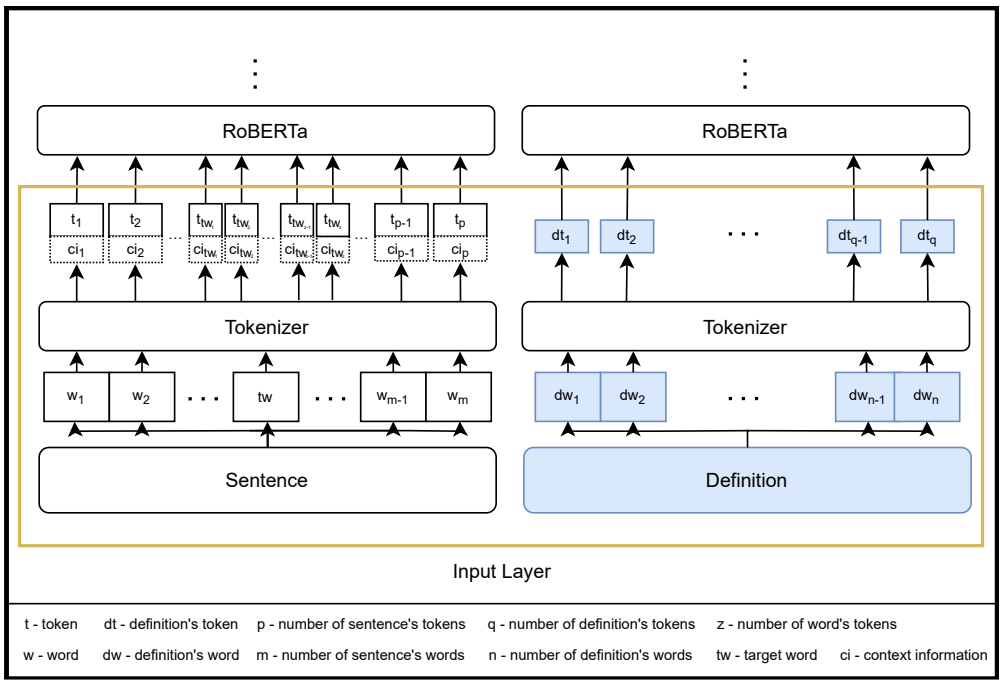

FIGURE 4.2: Input layer using RoBERTa. As single words are often divided into multiple tokens, most of the times the number of input words does not equal the number of output tokens. Context information is used to determine all the tokens that correspond to the target word.

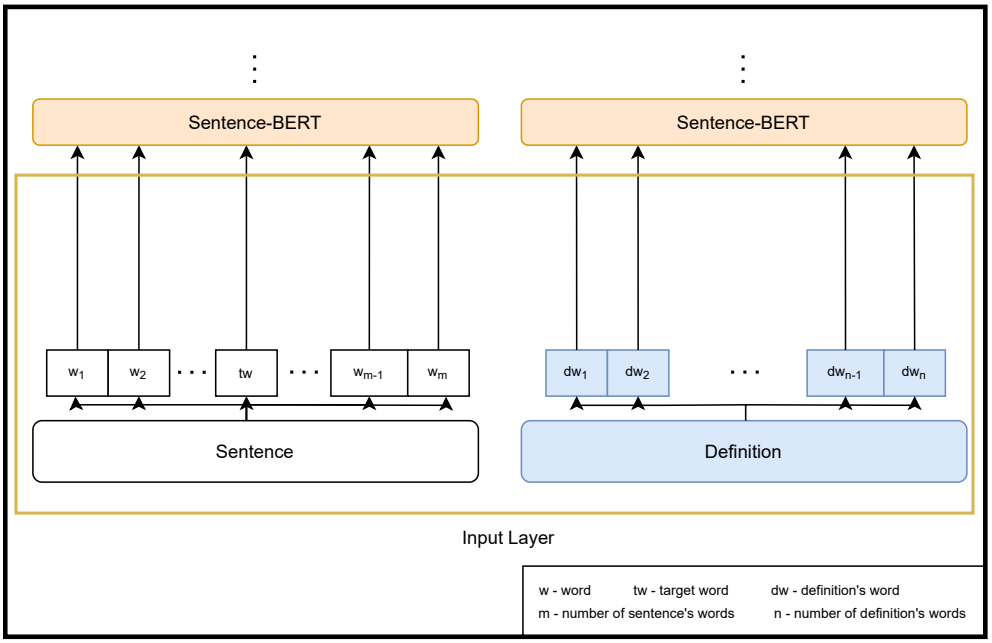

FIGURE 4.3: Input layer using Sentence-BERT. In one of the sub-models, MIss WiLDe takes advantage of Sentence-BERT instead of RoBERTa.

the embedding representation of the sentence is constructed. In the work of Choi et al., sentence representation is calculated using the [CLS] special token. However, it has been experimentally established by Reimers and Gurevych in their work on Sentence-BERT [89] that this approach falls short in comparison to using the mean value of tokens without the [CLS] and [SEP] tokens. It was also further confirmed by the results I achieved in a number of experimental trials with and without the use of the aforementioned special tokens.

I present three variants of the MIss WiLDe model, which I am going to interchangeably call the sub-models. These are:

- MIss WiLDe_base. This is the core version of my model. See Figure 4.1 for the model overview and Figure 4.2 for its input layer using RoBERTa;
- MIss WiLDe_cos. Both SPV and MIP are methods of using semantic gaps to determine if a target word is used metaphorically. Therefore, I also created a sub-model using cosine similarity to explicitly handle semantic gaps. This is shown by CS in Figure 4.1. Specifically, similarity between the meaning of the sentence and the meaning of the target word is calculated within the SPV block, while similarity between the meaning of the target word's definition and the meaning of the target word itself is calculated within MIP block. The input layer for this sub-model is common with the base variant visualized in Figure 4.2;

- MIss WiLDe_sbert. Since the results published in [89] suggest that using Sentence-BERT should result in acquiring sentence embeddings of better quality than those produced by both [CLS] tokens and averaged token vectors, I have decided to confirm it experimentally. I have therefore replaced RoBERTa with Sentence-BERT[2] as an encoder in one of my three sub-models. The input layer using Sentence-BERT is depicted in Figure 4.3.

The input to my model consists of the sentence comprising the target word on one side, and the definition of this target word on the other (depending on the part of speech and the availability of the definition in Wiktionary, lemma can be used instead of the definition; cf. Section 4.5 for the details). The conversion of words into tokens is then performed using the improved implementation of Byte-Pair Encoding (BPE), as proposed by Radford et al. in [87] and used by Choi et al. in [16] as well. This can be described as follows:

- $TOK(w_1, w_2, \cdots tw, \cdots, w_{m-1}, w_m) = t_1, t_2, \cdots t_{tw}, \cdots, t_{p-1}, t_p$;
- $TOK(dw_1, dw_2, \cdots, dw_{n-1}, dw_n) = dt_1, dt_2, \cdots, dt_{q-1}, dt_q$

where TOK stands for the tokenizer, w represents a single word within the analyzed sentence, with tw being the target word or, to put it differently, a metaphor candidate; m in the subscript is the number of words in the input sentence. t represents an output token, while p is the number of the output tokens. Depending on a given target word tw, t_{tw} should be

[2]Specifically, SentenceTransformer("roberta-base-nli-stsb-mean-tokens") was utilized.

considered an abbreviation for $t_{tw_1}, t_{tw_2}, \cdots, t_{tw_{z-1}}, t_{tw_z}$, where z stands for the number of tokens the target word was split into. This can be observed in Figure 4.2 as well. In the formulas presented in this section, I use the abbreviated forms for simplicity[3]. In the second formula, dw stands for a component word of the target word's definition and dt for an output token; n and q in the subscripts denote the number of words in the definition and the number of related output tokens, respectively.

Afterwards, tokens are transformed into embedding vectors via the encoder layer. Input and output of the two encoders can be illustrated with the following formulas:

- $ENC(t_1, t_2, \cdots t_{tw}, \cdots, t_{p-1}, t_p) = te_1, te_2, \cdots te_{tw}, \cdots, te_{p-1}, te_p$;
- $ENC(dt_1, dt_2, \cdots, dt_{q-1}, dt_q) = dte_1, dte_2, \cdots, dte_{q-1}, dte_q$

where ENC stands for the function producing contextualized vector representation for a given input, t represents a single token within the analyzed sentence, with tw in the subscript denoting the target word. te is the vector embedding representation that corresponds to the input token with the same index, while te_{tw} is the embedding representation of the target word's tokens. Analogously, dt stands for a token coming from the definition of the target word and dte for a vector embedding of the said token. Additionally, p and q denote the length of the sentence and the length of the target word's definition measured in the number of tokens, respectively.

[3]A single input word is often transformed into multiple tokens: for more details on Byte-Pair Encoding cf. https://huggingface.co/docs/transformers/tokenizer_summary; last accessed on 1 March 2022.

Subsequently, the mean value of the sentence's token vectors on one side and the mean value of the definition's token vectors on the other are computed within the pooling layer. To the output of this layer, dropout is subsequently applied. On both sides, an output vector is then concatenated with the vector representation of the target word's tokens having undergone the same operations. In case of the cosine-similarity sub-model, an additional third vector representing similarity between the two respective vectors is also concatenated. In order to calculate the gap between these vectors, a multilayer perceptron is applied on the output of the concatenation function. The formulas for the hidden vectors obtained this way in SPV and MIP layers are presented below.

- $hv_{spv} = \Phi(\frac{1}{p}\sum_{i=1}^{p} te_i, te_{tw})$;
- $hv_{mip} = \Phi(\frac{1}{q}\sum_{j=1}^{q} dte_j, te_{tw})$

where Φ represents concatenation; p and q denote length of the sentence and length of the definition, respectively (measured in number of tokens); i is the index of the sentence's token such that $i \in \mathbb{Z}$, $p-1 \geq i \geq 1$; and j is the index of the definition's token such that $j \in \mathbb{Z}$, $q-1 \geq j \geq 1$. The hidden vector hv_{spv} represents a vector being the output of the SPV layer while the hv_{mip} is used to depict a hidden vector that is the output of the MIP layer.

As mentioned, for the cosine-similarity sub-model, the similarity vector becomes the third element concatenated in order

to obtain the aforementioned hidden vectors. Similarity vectors are obtained as follows:

- $similarity_{spv} = \frac{\frac{1}{p}\sum_{i=1}^{p} te_i \cdot te_{tw}}{max\left(\left\|\frac{1}{p}\sum_{i=1}^{p} te_i\right\|_2 \cdot \|te_{tw}\|_2, \varepsilon\right)}$;
- $similarity_{mip} = \frac{\frac{1}{q}\sum_{j=1}^{q} dte_j \cdot te_{tw}}{max\left(\left\|\frac{1}{q}\sum_{j=1}^{q} dte_j\right\|_2 \cdot \|te_{tw}\|_2, \varepsilon\right)}$

where $similarity_{spv}$ stands for the cosine similarity value measured between the average sentence vector and the target word vector; analogously $similarity_{mip}$ represents the cosine similarity between the average definition vector and the target word vector. The $\|\|_2$ denotes the Euclidean norm, the $\cdot$ stands for the dot product between the vectors, and ε is a parameter of small value used to avoid division by zero[3]. The output of the model is calculated by adding bias to the concatenation function that takes in the two hidden vectors and applies the log-softmax activation function onto the result. Finally, the candidate with the higher probability score is chosen as the predicted label. The process can be represented with the following formula:

- $y_\tau = \log(\sigma(W^\top \Phi(hv_{spv}, hv_{mip}) + b))$;
- $\hat{y} = \text{argmax}(y_\tau)$

where $\hat{y}$ is the label predicted by the model, such that $\hat{y} \in \{0, 1\}$. This prediction is the result of the argmax operation applied on y_τ, which in turn stands for the natural logarithm of the value output by the softmax function denoted with σ. Softmax outputs two values that are the probabilities for

[3] https://pytorch.org/docs/stable/generated/torch.nn.CosineSimilarity.html; last accessed on 1 March 2022.

each class (literal and metaphor) that range from 0 to 1 and sum up to 1. $W^\top$ denotes the weights matrix, Φ stands for concatenation, and b signifies bias.

I use the negative log-likelihood loss function, which in combination with log-softmax activation, acts essentially the same as cross-entropy combined with softmax, but has improved numerical stability in PyTorch [70].

Visualization of the model is provided in Figure 4.1. Two variants of its input layer can be compared as shown in Figure 4.2 and 4.3. The code allowing for training the three sub-models and reproducing the results can be found at my GitHub[4].

4.4 Datasets

In this section, I present the datasets used in the experiments. I wanted to confirm the validity of my hypothesis that utilizing lexical definitions of target words would improve the algorithm's performance and in order to do so I have adopted the same datasets as in the work of Choi et al. [16]. The original data is available at the authors' drive, a link to which can be found on their GitHub[5]. The downloadable repository consists of MOH-X, TroFi, VUA-20 (variant of VUA-ALL-POS

[4] https://github.com/languagemedialaboratory/ms_wilde; last updated on 17 February 2022.

[5] https://drive.google.com/file/d/1738aqFObjfcOg2O7knrELmUHulNhoqRz/view?usp=sharing via https://github.com/jin530/MelBERT; last accessed on 1 March 2022.

known from Metaphor Detection Shared Task [61, 60]), VUA-18 (variant of VUA-SEQ known from Gao et al. [30] and Mao et al. [66]), VUA-VERB, eight subsets of VUA-18, four of which are selected based on the POS tags of the target words (nouns, verbs, adjectives, and adverbs), and another four on the genre to which the sentence comprising target word belongs (academic, conversation, fiction, and news). Both Genres and POS are used only for testing. The same datasets enriched with the Wiktionary definitions can be downloaded directly from my GitHub[6].

MOH-X (Mohammad et al. [2016] dataset) [72] and TroFi (Trope Finder) [8] are relatively small datasets annotated only for verbs. MOH-X is built with the example sentences taken from WordNet [28], and TroFi with the ones from the Wall Street Journal [14]. For the sake of fair comparison with Choi et al. [16], I use these two datasets only as the test sets for the models trained beforehand on the VUA-20, in the same way as performed by the authors of MelBERT. As they note in the paper, this can be viewed as the zero-shot transfer learning.

VUAMC (Vrjie Universiteit Amsterdam Metaphor Corpus)[7] [100, 98] is the biggest publicly available corpus annotated for token-level metaphor detection and it is seemingly the most popular one in the field. It comprises text

[6] https://github.com/languagemedialaboratory/ms_wilde/tree/main/data; last updated on 17 February 2022.

[7] http://www.vismet.org/metcor/documentation/home.html; last accessed on 1 March 2022.

fragments sampled from the British National Corpus[8]. Sentences it contains were labeled in accordance with MIPVU (Metaphor Identification Procedure VU University Amsterdam), the refined and adjusted version of the already described MIP (Metaphor Identification Precedure). Both VUA-ALL-POS (All Parts of Speech) and VUA-SEQ (Sequential) are based on VUAMC, which has been used in the Metaphor Detection Shared Task, first in 2018 and later in 2020 [61, 60].

The repository prepared for the Metaphor Detection Shared Task is provided on the organizer's GitHub[9]. Inside, one can find the links allowing for downloading:

- VUAMC corpus in XML format;
- Starter kits for obtaining training and testing splits of VUAMC corpus (vuamc_corpus_train, vuamc_corpus_test);
- Lists of ids (all_pos_tokens, all_pos_tokens_test, verb_tokens, verb_tokens_test) specifying the tokens from VUAMC to be used as targets for classification in the two tracks of the Metaphor Detection Shared Task: All-Part-Of-Speech and Verbs.

In the 12,122 sentences comprised by vuamc_corpus_train, all of their component words are labeled for metaphoricity, irrespective of the part of speech they belong to. For example, in the input sentence "The villagers seemed unimpressed , but

[8] http://www.natcorp.ox.ac.uk/; last accessed on 1 March 2022.

[9] https://github.com/EducationalTestingService/metaphor/tree/master/VUA-shared-task; last accessed on 1 March 2022.

were M_given no choice M_in the matter .", there are altogether 14 tokens, including punctuation marks. Two of the tokens are labeled as metaphors (the verb *given* and the preposition *in*), and the remaining twelve as non-metaphors. This is indicated by the prefix "M" attached to the metaphors. In all_pos_tokens, only six out of these 14 tokens are chosen as targets for classification. These tokens are: *villagers*, *seemed*, *unimpressed*, *given*, *choice*, and *matter*. In this dissertation, after Neidlein et al. [73], the name VUA-ALL-POS refers to the dataset utilizing only the target words specified by all_pos_tokens and all_pos_tokens_test. The dataset called VUA-20, which I adopt from Choi et al. [16] and which I use in the experiments, comprises the same testing data, yet it produces more training samples from vuamc_corpus_train than specified by all_pos_tokens. In the example sentence above, VUA-20 uses all of the available tokens, excluding punctuation, as targets for classification. VUA-20 takes advantage of both content words (belonging to verbs, nouns, adjectives and adverbs) and function words (members of remaining parts of speech), while VUA-ALL-POS is said to limit itself to the content words only (excluding verbs *have*, *do* and *be*). This difference results in a much bigger number of target tokens available in the former's training set (160,154 and 72,611 for VUA-20 and VUA-ALL-POS, respectively). At the same time, for reasons unknown, VUA-20 lacks 86 of the target tokens used in VUA-ALL-POS. With

this exception, VUA-20 can be therefore viewed as an extended variant of VUA-ALL-POS. VUA-20 includes all of the sentences utilized by VUA-ALL-POS plus those excluded from the latter due to the POS-related restrictions. As a result, while there are 12,122 unique sentences provided in total by vuamc_corpus_train, the numbers of sentences used for training in VUA-20 and VUA-ALL-POS are 12,093 and 10,894, respectively. The 29 "sentences", which VUA-20 is lacking with respect to vuamc_corpus_train, were excluded, presumably because they are either empty strings or single punctuation characters (“”, “.”, “!”, and “?”). As mentioned, testing data is common for both datasets: they comprise the same 22,196 target tokens coming from 3,698 sentences selected from 4,080 available in total in vuamc_corpus_test.

VUA-SEQ (I use this name after Neidlein et al. [73]) is another dataset built upon VUAMC. It was used in the works of Gao et al. [30] and Mao et al. [66], among the others. It differs from VUA-ALL-POS in that it employs different splits of VUAMC and in that it uses all of the tokens available in a sentence as targets for classification (including punctuation marks). This results in a much bigger number of target tokens used by VUA-SEQ in comparison with VUA-ALL-POS (205,425 and 94,807, respectively). However, VUA-SEQ uses a smaller number of unique sentences than VUA-ALL-POS (10,567 and 14,974, respectively). Unlike VUA-ALL-POS, VUA-SEQ has a development set as well. VUA-18, which I adopt from Choi et al. [16], is very similar to VUA-SEQ, as it

uses the same sentences in each of the subsets (6,323, 1,550, and 2,694 sentences for training, development, and testing sets, respectively). What does not allow for calling the two datasets identical is that VUA-18 does not count contractions and punctuation marks as separate tokens (there is a very small number of exceptions to this general rule). For example, the sentence coded in VUAMC as: "M_Lot of M_things daddy has n't seen ." is divided into eight tokens in VUA-SEQ, whereas in VUA-18 it is presented as "Lot of things daddy hasn't seen.", which results in using only six tokens and six corresponding labels (without "n't" and "."). In consequence, the numbers of tokens are: 116,622 and 101,975 in the training sets, 38,628 and 34,253 in the development sets, 50,175 and 43,947 in the testing sets of VUA-SEQ and VUA-18, respectively. To stay in uniformity with Choi et al. [16], I am not utilizing VUA-18's development set in the experimental trials.

VUA-VERB, which I adopt from Choi et al. [16], utilizes the same sentences as those selected in the lists prepared for the Metaphor Detection Shared Task (verb_tokens and verb_tokens_test), although it splits the original training data into training and validation subsets. While in verb_tokens there are 17,240 target tokens used for training, in VUA-VERB there are 15,516 and 1,724 tokens comprised by its training and development sets, respectively. The number of tokens used for testing equals 5,873, which is the same in both cases. In the experimental trials, I am not taking advantage

of VUA-VERB's development set.

Although Neidlein et al. [73] claim that it is only the content words (verbs, nouns, adjectives, and adverbs), whose labels are being predicted in VUA-ALL-POS, this is not entirely accurate. There are instances of interjections (*Ah* in "Ah , yeah ."), prepositions (*like* in "(...) it would be interesting to know what he thought children were like ."), conjunctions (*either* in "(...) criminal behaviour is either inherited or a consequence of unique , individual experiences ."), etc.

While in their paper Su et al. [104, p. 32] formulate the opinion that "POS such as punctuation, prepositions, and conjunctions are unlikely to trigger metaphors", at the same time they provide the evidence for the opposite, at least with regard to the prepositions: from Figure 5 they attach [104, p. 35], it is clear that adpositions are annotated as being used metaphorically more often than any other part of speech in VUAMC. This should come as no surprise, as, for example, there are entire tomes devoted to the analysis of the primarily spatial senses of temporal prepositions and related metaphorical meaning extensions (cf. [10, 41, 85]).

4.5 Data Preprocessing

For data preprocessing, I use Python equipped with WordNetLemmatizer[11] for obtaining the dictionary forms of given

[11] https://www.nltk.org/_modules/nltk/stem/wordnet.html; last accessed on 1 March 2022.

words and Wiktionary Parser[11] for retrieving their definitions. As the outputs are different for different parts of speech, I take advantage of the POS tags already included in the datasets. These, however, have to be first mapped to the format used by Wiktionary. It is noteworthy that not all of these POS tags are accurate, which sometimes prevents the algorithm from retrieving the appropriate definition.

In Delahunty and Garvey's book [20], nouns, verbs, adjectives, and adverbs are termed as the *major parts of speech*, and as such are contrasted with the *minor parts of speech* (all the others). Alternatively, they can be called the *content words* and the *function words*, respectively (cf. Haspelmath's analysis [40]). The former have more specific or detailed semantic content, while the latter have a more non-conceptual meaning and fulfill an essentially grammatical function [18]. As a result of this, in general, definitions of the function words do not add much of semantic information that could be used in metaphor detection. On the contrary, using averaged vectors of their component tokens could become a source of unnecessary noise, leading to performance decay (consider the first definition of the very frequently occurring determiner *the* found in Wiktionary: "Definite grammatical article that implies necessarily that an entity it articulates is presupposed; something already mentioned, or completely specified later

[11] https://github.com/Suyash458/WiktionaryParser; last accessed on 1 March 2022.

in that same sentence, or assumed already completely specified"). When it comes to function words, I therefore decided to use their lemmas instead of definitions. In this regard, I have made an exception for twelve prepositions, namely: *above*, *below*, *between*, *down*, *in*, *into*, *on*, *over*, *out*, *through*, *under*, and *up*. These are all very frequent, as they all constitute a part of the stopwords from NLTK's *corpus* package[12]. When present in utterances, they often manifest the underlying image schemata well known from the Conceptual Metaphor Theory first popularized by Lakoff & Johnson in [56] and further described in detail by other authors, for example by Gibbs in [33]. Admittedly, the choice of the words from the outside of the major parts of speech category is subjective and could be made differently.

I assumed that it is likely that the first definition available in the Wiktionary is going to be the one representing a word's *basic meaning* and therefore I retrieve only the first of the definitions available for the given part of speech. This choice was preceded with the lecture of Wiktionary guidelines[13], where — at least for the complex entries — it is explicitly recommended to use the logical hierarchy of the word senses, meaning: *core sense at the root*. The relation between basic meanings and metaphoricity within the logical hierarchy is explained in [102, p. 285] (emphasis added):

[12] https://www.nltk.org/_modules/nltk/corpus.html; last accessed on 1 March 2022.

[13] https://en.wiktionary.org/wiki/Wiktionary:Style_guide#Definitions; last accessed on 1 March 2022.

> The logical ordering runs from core senses to subsenses. *Core meanings* or *basic meanings* are the meanings which are felt as *the most literal* or central ones. The relation between core sense and subsense may be understood in various ways, e.g., as the relation between general and specialised meaning, central and peripheral, *literal and non-literal*, concrete and abstract, original and derived.

Following the general strategy of using only the first of the available definitions, I make an exception for the words whose first definitions include the tags *archaic* and *obsolete*. Although, as mentioned in Section 4.2.1, MIP (Metaphor Identification Procedure) postulates considering the historical antecedence as one of the cues in establishing a given word's basic meaning, studying the data coming from Wiktionary led me to the conclusion that very often the definitions comprising the aforementioned labels stand for the senses that are no longer accessible to contemporary language users. For this reason, I argue that they should not be treated as the *basic senses* and that it is not appropriate to compare them with the contextual senses in order to decide whether a word is used figuratively. In practice, my algorithm collects the first definition of a given target word without the words *archaic* and *obsolete* (or their derivatives) inside the brackets. For example, the definitions of the verb *consist* available in the Wiktionary are as follows:

1. (obsolete, copulative) To be.
2. (obsolete, intransitive) To exist.
3. (intransitive, with in) To be comprised or contained.

4. (intransitive, with of) To be composed, formed, or made up (of).

Out of these four, it is the third definition that includes neither of the tags mentioned above and thus becomes accepted by my algorithm. Furthermore, as with all other definitions I collect, the brackets along with their content are erased. The final shape of the definition adopted for the target word *consist* is therefore: *To be comprised or contained.* In the case where all the definitions of a given word include either *archaic* or *obsolete* labels, I keep the first definition from the list.

Using the algorithm illustrated above, the definitions are collected automatically and without supervision, which significantly reduces costs. Experimental results presented in the following section prove that this simple method is in fact efficient.

4.6 Experiments

In this section, I first present the models whose results are used as the baselines for comparison in Section 4.7. Subsequently, I provide a brief description of the setup used throughout the experiments.

4.6.1 Models for Comparison

I compare the performance of MIss WiLDe's three variants with nine other models, which are the following:

- **MDGI-Joint-S** and **MDGI-Joint** (denoted as MDGI-J-S and MDGI-J). These are the two variants of the model designed by Wan et al. in [109]. The outline of the model has already been presented in Section 2.3.2. The first variant of the model (MDGI-Joint-S) shares parameters between the context encoder and the definition encoder, while the other one uses independent encoders. Although in the paper authors present part of their results as achieved using VUA-ALL-POS, this seems to be inaccurate[15], and therefore I place them in Table 4.2, which shows the results for VUA-SEQ/VUA-18. The results for VUA-VERB reported by the authors can be found in Table 4.3. As for TroFi, I use it only as the test set for the model trained on VUA-20 and thus the results reported in [109] are not comparable with mine.
- **BERT**. The model presented by Neidlein et al. [73] using the uncased base BERT model as its backbone and some standard hyperparameters. In the following tables, I present the results reported by the authors.
- **RNN_HG** and **RNN_MHCA**. The two models built upon Gao et al. [30] and presented by Mao et al. in [66]. The first model follows the guidelines of MIP (Metaphor Identification Procedure), while the other follows SPV (Selectional Preference Violation). The first model uses GloVe (Global Vectors) embedding as the representation

[15]See the datasets available at the authors' GitHub: https://github.com/sysulic/MDGI; accessed on 21 December 2021.

of the target word's literal meaning and the hidden state from the BiLSTM fed with the concatenation of GloVe and ELMo embeddings as the representation of its contextual meaning. In order to compute the contextual representation of the target word, RNN_MHCA uses multi-head contextual attention.

- **RoBERTa_BASE** (denoted as RoB_BASE). A vanilla version of the RoBERTa model designed for metaphor detection prepared as a baseline by Choi et al. [16] Unlike MelBERT, it utilizes all-to-all interaction architecture.
- **RoBERTa_SEQ** (denoted as RoB_SEQ). To paraphrase Choi et al. [16], this model takes one single sentence as an input, where the target word is marked as the input embedding token. The model predicts the metaphoricity of the target word using its embedding vector.
- **DeepMet**. The model designed by Su et al. [104], the winner of the Metaphor Detection Shared Task 2020. Its outline has already been presented in Section 2.3.2.
- **MelBERT**. Model presented by Choi et al. in [16]. Its outline has already been presented in Section 2.3.2 and, in comparison with my model, in Section 4.3.
- **MIss WiLDe_base**, **MIss WiLDe_cos** and **MIss WiLDe_sbert** (denoted as MsW_base, MsW_cos, and MsW_sbert). These are the three variants of my model described in details in Section 4.3. As signalized by their names, the first one is a core model, the second one uses

a cosine-similarity measure as an additional feature, and the third one uses the Sentence-BERT encoder in place of RoBERTa, which is utilized in the first two sub-models.

4.6.2 Experimental Setup

In order to confirm my suppositions concerning possible improvements acquired through the introduction of the innovations outlined above, the experimental setup is kept the same as in the work by Choi et al. [16]. For the sake of brevity, I compare the models without using a bootstrapping aggregation technique. I use the same hyperparameters for training the models: The batch size is set to 32, the max sequence length is set to 150, and the number of epochs is set to 3. The AdamW optimizer is used with the initial learning rate of 3×10^{-5} and the third epoch is set as the warm-up epoch. I perform five trials for every experiment. The results presented in the tables are calculated, taking the mean value of the scores achieved over five runs. To preserve reproducibility and exclude any form of cherry-picking, I use the same set of random seeds: 1, 2, 3, 4, and 5. The same GPU, specifically Tesla P100-PCIE-16GB provided by Google Colab[15], is used for every experiment performed. The code is implemented using PyTorch.

[15] https://colab.research.google.com/; last accessed on 1 March 2022.

4.7 Results

In this section I present the results achieved on the datasets described in Section 4.4 and compare them with the models introduced in Section 4.6.1. If in Section 4.6.1 it is not stated otherwise, the results in the tables yielded by the models other than the three variants of MIss WiLDe (my proposed method) are cited from Choi et al. [16]. The bold font and the underline stand for the best and the second-best results, respectively. Choi et al. [16] report the results up to one place after the decimal point, which sometimes did not allow for a definite assertion as to which model performed better. As a consequence, there are columns with two values marked in bold or underlined. Using a horizontal line, I separate the results achieved by my proposed method from those achieved by other models.

TABLE 4.1: Results for VUA-ALL-POS (All Parts of Speech) and VUA-20 (variant of VUA-ALL-POS from Choi et al.; cf. Section 4.4 for details).

Dataset	Model	Precision	Recall	F1
VUA-ALL-POS	BERT	75.4%	64.7%	69.7%
VUA-20	RoB_BASE	71.7%	60.2%	65.5%
	RoB_SEQ	**76.9%**	66.7%	71.4%
	DeepMet	<u>76.7%</u>	65.9%	70.9%
	MelBERT	76.4%	68.6%	72.3%
	MsW_base	75.4%	<u>69.9%</u>	<u>72.5%</u>
	MsW_cos	75.4%	**70.3%**	**72.7%**
	MsW_sbert	76.2%	66.5%	71.0%

TABLE 4.2: Results for VUA-SEQ (Sequential) and VUA-18 (variant of VUA-SEQ from Choi et al.; cf. Section 4.4 for details).

Dataset	Model	Precision	Recall	F1
VUA-SEQ	MDGI-J-S	81.3%	73.2%	77.0%
	MDGI-J	**82.5%**	72.5%	77.2%
	BERT	78.0%	76.9%	77.5%
VUA-18	RNN_HG	71.8%	76.3%	74.0%
	RNN_MHCA	73.0%	75.7%	74.3%
	RoB_BASE	79.4%	75.0%	77.1%
	RoB_SEQ	80.4%	74.9%	77.5%
	DeepMet	82.0%	71.3%	76.3%
	MelBERT	80.1%	76.9%	78.5%
	MsW_base	79.6%	78.3%	**78.9%**
	MsW_cos	79.3%	**78.5%**	78.9%
	MsW_sbert	75.6%	71.0%	72.7%

TABLE 4.3: Results for VUA-VERB.

Dataset	Model	Precision	Recall	F1
VUA-VERB	MDGI-J-S	78.8%	71.5%	75.0%
	MDGI-J	78.9%	70.9%	74.7%
	RNN_HG	69.3%	72.3%	70.8%
	RNN_MHCA	66.3%	75.2%	70.5%
	RoB_BASE	76.9%	72.8%	74.7%
	RoB_SEQ	79.2%	69.8%	74.2%
	DeepMet	**79.5%**	70.8%	74.9%
	MelBERT	78.7%	72.9%	**75.7%**
	MsW_base	62.1%	77.1%	68.8%
	MsW_cos	60.9%	**77.7%**	68.3%
	MsW_sbert	67.2%	77.7%	72.0%

As shown in Tables 4.1 and 4.2, when using either VUA-20 or VUA-18, MIss WiLDe manages to outperform all other models in terms of F1-score and recall. It achieves the best recall on VUA-VERB as well, however performs significantly

TABLE 4.4: Results for Genres.

Genre	Model	Precision	Recall	F1
Acad	RNN_HG	76.5%	83.0%	79.6%
	RNN_MHCA	79.6%	80.0%	79.8%
	RoB_BASE	88.1%	79.5%	83.6%
	RoB_SEQ	86.0%	77.3%	81.4%
	DeepMet	**88.4%**	74.7%	81.0%
	MelBERT	85.3%	82.5%	83.9%
	MsW_base	85.3%	82.4%	83.8%
	MsW_cos	84.4%	**83.5%**	**84.0%**
	MsW_sbert	85.6%	81.0%	83.2%
Conv	RNN_HG	63.6%	72.5%	67.8%
	RNN_MHCA	64.0%	71.1%	67.4%
	RoB_BASE	70.3%	69.0%	69.6%
	RoB_SEQ	70.5%	69.8%	70.1%
	DeepMet	**71.6%**	71.1%	**71.4%**
	MelBERT	70.1%	71.7%	70.9%
	MsW_base	69.2%	**73.8%**	71.4%
	MsW_cos	70.2%	72.7%	**71.4%**
	MsW_sbert	68.0%	73.5%	70.6%
Fict	RNN_HG	61.8%	74.5%	67.5%
	RNN_MHCA	64.8%	70.9%	67.7%
	RoB_BASE	74.3%	72.1%	73.2%
	RoB_SEQ	73.9%	72.7%	73.3%
	DeepMet	**76.1%**	70.1%	73.0%
	MelBERT	74.0%	76.8%	75.4%
	MsW_base	73.6%	**77.6%**	**75.5%**
	MsW_cos	73.3%	77.1%	75.2%
	MsW_sbert	72.9%	77.3%	75.0%
News	RNN_HG	71.6%	**76.8%**	74.1%
	RNN_MHCA	74.8%	75.3%	75.0%
	RoB_BASE	83.5%	71.8%	77.2%
	RoB_SEQ	82.2%	74.1%	77.9%
	DeepMet	**84.1%**	67.6%	75.0%
	MelBERT	81.0%	73.7%	77.2%
	MsW_base	82.2%	76.1%	**79.0%**
	MsW_cos	81.5%	76.0%	78.7%
	MsW_sbert	81.6%	74.3%	77.8%

worse in regards to both the precision and F1-score. Surprisingly, in other cases an overall weaker variant of my model

based on Sentence-BERT yields the best scores out of the three variants. The results for VUA-VERB are compared in Table 4.3. Tables 4.4 and 4.5 show the results for Genres and Parts of Speech, respectively. As for Genres, at least one of my sub-models manages to either outperform all other models or at least to draw with one of the competitors. Again MIss WiLDe performs overall better than the other models in terms of recall, while it loses in precision. When it comes to Parts of Speech, specifically verbs and adjectives, the base variant of MIss WiLDe proves to be by far the best model in terms of F1-score. This raises the question as to why it is not performing similarly well on the VUA-VERB dataset. The answer might be that for Genres and POS, the results are presented for the models trained beforehand on VUA-18, which provides a significantly larger training set then VUA-VERB. Lastly, Tables 4.6 and 4.7 show the results for MOH-X and TroFi. As described in Section 4.4, following the approach of Choi et al. [16], these datasets are used only for testing with the models trained beforehand on VUA-20. When using MOH-X, the base variant of my model outperforms the competition in terms of F1-score. Neither of my models manages to win on TroFi; it is the model of Choi et al. [16] that achieves the best F1-score and recall for the said dataset.

Although, as mentioned before, in the tables I quote the results published by Choi et al. in [16], at the same time I would like to share the results of MelBERT's reruns for both VUA-20 and VUA-18 that I have conducted myself using the same set

TABLE 4.5: Results for POS (Parts of Speech).

POS	Model	Precision	Recall	F1
Verb	RNN_HG	66.4%	75.5%	70.7%
	RNN_MHCA	66.0%	76.0%	70.7%
	RoB_BASE	77.0%	72.1%	74.5%
	RoB_SEQ	74.4%	75.1%	74.8%
	DeepMet	**78.8%**	68.5%	73.3%
	MelBERT	74.2%	75.9%	75.1%
	MsW_base	74.9%	**78.1%**	**76.5%**
	MsW_cos	74.0%	77.9%	75.9%
	MsW_sbert	74.5%	76.1%	75.3%
Adj	RNN_HG	59.2%	**65.6%**	62.2%
	RNN_MHCA	61.4%	61.1%	61.6%
	RoB_BASE	71.7%	59.0%	64.7%
	RoB_SEQ	72.0%	57.1%	63.7%
	DeepMet	**79.0%**	52.9%	63.3%
	MelBERT	69.4%	60.1%	64.4%
	MsW_base	68.5%	65.0%	**66.7%**
	MsW_cos	67.5%	65.1%	66.3%
	MsW_sbert	69.3%	63.8%	66.4%
Adv	RNN_HG	61.0%	66.8%	63.8%
	RNN_MHCA	66.1%	60.7%	63.2%
	RoB_BASE	78.2%	69.3%	73.5%
	RoB_SEQ	77.6%	63.9%	70.1%
	DeepMet	79.4%	66.4%	72.3%
	MelBERT	**80.2%**	69.7%	**74.6%**
	MsW_base	76.4%	69.3%	72.7%
	MsW_cos	78.9%	**69.8%**	74.0%
	MsW_sbert	78.3%	67.5%	72.4%
Noun	RNN_HG	60.3%	**66.8%**	63.4%
	RNN_MHCA	69.1%	58.2%	63.2%
	RoB_BASE	**77.5%**	60.4%	67.9%
	RoB_SEQ	76.5%	59.0%	66.6%
	DeepMet	76.5%	57.1%	65.4%
	MelBERT	75.4%	66.5%	**70.7%**
	MsW_base	76.2%	64.9%	70.1%
	MsW_cos	75.4%	66.0%	70.4%
	MsW_sbert	73.9%	64.7%	69.0%

TABLE 4.6: Results for MOH-X (Mohammad et al. [2016] dataset).

Dataset	Model	Precision	Recall	F1
MOH-X	DeepMet	79.9%	76.5%	77.9%
	MelBERT	79.3%	79.7%	79.2%
	MsW_base	82.3%	**80.0%**	**80.8%**
	MsW_cos	81.0%	80.0%	80.2%
	MsW_sbert	**83.3%**	77.7%	80.1%

TABLE 4.7: Results for TroFi (Trope Finder).

Dataset	Model	Precision	Recall	F1
TroFi	DeepMet	**53.7%**	72.9%	61.7%
	MelBERT	53.4%	**74.1%**	**62.0%**
	MsW_base	53.3%	73.4%	61.7%
	MsW_cos	53.2%	72.8%	61.4%
	MsW_sbert	**53.7%**	69.7%	60.6%

of random seeds as well as the same GPU. These are as follows: VUA-20 (precision: 75.8%, recall: 69.6%, F1: 72.6%); VUA-18 (precision: 79.9%, recall: 77.3%, F1: 78.6%). Additionally, I have compared these F1-scores with the F1-scores achieved by MsW_base and MsW_cos, using two-tailed t-test. The value achieved was $p > 0.05$, meaning that the differences are not statistically significant.

4.8 Considerations

In this section I present my considerations regarding the experimental results introduced above. Example sentences presented in this section come from the output of the experiment

performed on VUA-20 sharing common test data with VUA-ALL-POS which is considered one of the most representative datasets for evaluating the performance of the algorithms designed for metaphor detection. Target words are marked in bold font. I compare the predictions of Choi et al.'s MelBERT [16] and my model. Out of the three variants of MIss WiLDe, I have chosen the base one. In doing so, I can argue that the differences in the output stem mostly from my decision to use lexical definitions instead of the isolated target word. As mentioned earlier, for each of my models I have conducted five runs of experiments, using the same random seeds ranging from 1 to 5 for fairness. Since the output and its specific values may slightly vary between random seeds, for both MelBERT and Miss WiLDe_base, I use the outputs of the models having yielded the best F1-score among all five seeds.

4.8.1 Results Analysis

As can be told from the tables above, MIss WiLDe managed to outperform competing models in several categories, most significantly in adjectives (cf. Table 4.5). Consider the following example:

> *Eventually they will be replaced, but more than 60 years on they run with the* ***rhythmic*** *reliability of a Swiss watch.*

Here my model voted for metaphorical use with a high degree of confidence (0.2:0.8), while MelBERT estimated it was literal (0.66:0.34). The definition of the target word retrieved

from Wiktionary and used by my model is 'Of or relating to rhythm.' and this of the modified noun is 'The quality of being reliable, dependable, or trustworthy.' In terms of Wilks [111], MIss WiLDe managed to detect the violation of semantic restrictions present between basic senses of the target adjective *rhythmic* and the noun *reliability*.

Another example including an adjective is the following:

> *There is a* ***refreshing*** *simplicity and tenderness in Motion's account of the way Francis nurses her, but she herself is too sketchily drawn for the episode to carry much weights.*

The definition of *refreshing* is 'That refreshes someone; pleasantly fresh and different; granting vitality and energy', while *simplicity* is defined as 'The state or quality of being simple'. Although the middle part of the adjective's definition seems to already point at the figurative meaning of the word, it can still be argued that overall it is *more* literal than metaphorical and therefore it constitutes the word's *basic meaning*. In the prototypical situation, for its modifier, *refreshing* would demand a noun belonging to the semantic field of FLUIDS, DRINKS. *Simplicity* is not fulfilling this condition, which hints at figurative usage. Although the phrase *refreshing simplicity* is quite commonly used, my model managed to detect that it is metaphorically motivated. Due to its high frequency in corpora, it would be difficult to argue that such word matching is unnatural or exotic. As recent language models are pretrained also on the texts from the books,

where juxtapositions similar to *refreshing simplicity* are observed quite often, without the use of definitions, it would be difficult for them to discern metaphoricity underlying such wordings. In other words, without the definitions conveying the *basic meanings* it would be hard to find any indication of breaking some semantic restrictions. In this case, MelBERT voted 0.84:0.16 for *refreshing* to be used literally, while MIss WiLDe leaned towards metaphoricity with the probabilities of 0.41:0.59. By looking at the estimation scores, it can be assumed that making the decision was not easy for my model, which can be explained with the reasoning just outlined.

> *An **incoming** Labour government would turn large areas of Whitehall upside down (…)*

Phrase *incoming Labour government* from the sentence above can be viewed as an example of TIME AS SPACE metaphor [27, 86], where the target word *incoming* is used in the sense of 'future'. Its primary meaning related to physical motion is described by the first definition found in Wiktionary and thus provided to my model ('Coming in; arriving'). The contextual meaning is portrayed by the second of the Wiktionary's definitions ('Succeeding to an office'). This example also shows that, as a general rule, my choice to use only the first definitions as the ones likely to convey basic meaning was right. In this example, MelBERT predicted *incoming* as being used literally (0.67:0.33), while MIss WiLDe made a correct and firm decision voting for a metaphor (0.14:0.86).

4.8.2 Error Analysis

It should be noted that a word's basic meaning sometimes does not provide much help in metaphor detection. Consider the following example:

> *There are always accusations of **piracy** and copy-catting, though they can't usually be substantiated.*

The target word's definition in this case is 'Robbery at sea, a violation of international law; taking a ship away from the control of those who are legally entitled to it', which indeed probably should be considered the word's *basic* sense, as it is historically older than 'The unauthorized duplication of goods protected by intellectual property law' (the third from the definitions listed in Wiktionary). One should notice that in this case the target word does not violate any selectional preferences. Both senses relate to the notion of CRIME and the target word used in either of them does not sound particularly awkward in the given context. In such cases, utilizing the information provided by the SPV (Selectional Preference Violation) module does not resolve the problem. As my method adopts only simplified interpretation of MIP (Metaphor Identification Procedure) as well, it cannot differentiate between the word senses based on historical reasons. When encountering samples of this kind, my model is rather powerless.

In my opinion, some of the missed predictions are the outcome of an inaccurate annotation process. Consider the following example:

***See** if you can rustle up a cup of tea for Paula and me, please.*

Although the target word from the above sentence is annotated as a non-metaphor, I strongly believe that in this context, it is used figuratively. Following the MIP procedure, the annotator should first read the whole sentence to understand its meaning. Next, it should be determined whether the target word has a more basic meaning in other contexts. The first Wiktionary definition of the target word is 'To perceive or detect someone or something with the eyes, or as if by sight'. At this point it should be already clear that in this case, the contextual meaning of the verb *see* differs from the meaning described by the first definition. The context suggests that semantically it is closer to 'To determine by trial or experiment; to find out (if or whether)', the eighth of the Wiktionary definitions for the verb *see*. As the basic meaning tends to be more concrete and related to bodily action [36, p. 4], it should be rather straightforward to judge the contextual meaning as non-basic[17]. Assuming that I am correct, I have to give credit to MelBERT, which estimated very confidently that it deals with a metaphor (0.04:0.96). MIss WiLDe also voted for a metaphor, however apparently it was not an easy choice (0.41:0.59).

[17]It might be, however, that the annotators decided to treat the expression as based on a metonymy rather than on a metaphor. For a related discussion, see [97, pp. 57–61].

Chapter 5

Conclusions and Future Work

In this dissertation, I presented the outcome of my research in automatic metaphor detection using machine learning.

In Chapter 3, I reported on the two sets of classification experiments conducted using the datasets in Japanese. The task was to predict whether an input sentence belongs to the non-literal or the literal dataset. Non-literal data was retrieved from the digitalized dictionary of figurative expressions [76], where all the sentences were selected by hand. As a source of literal data, I used a random mix of sentences acquired from Wikipedia articles, local parliamentary minutes, and news articles. Testing data comprised sentences coming from literary texts and labeled by three human annotators.

In the first trial, I was comparing the performance of feature vectors representing seven sets of seed words whose choice was inspired by the cognitive-linguistic approach to metaphor. These sets contained the words often employed in figurative language, belonging to categories such as colors, body parts, and spatial positions. The values of the aforementioned input vectors were expressing presence or absence of said words

in a target sentence. Depending on the combination of the categories used for prediction, F-1 score yielded by the proposed method exceeded 70%, which suggests that the selected features were useful.

In the second trial, each input sentence was represented by a word-count vector of length equal to the size of the vocabulary available in the training set after excluding stopwords and simile indicators (*yō*, *mitai*, *kurai* etc.). Using such features, I compared the performance of three popular classification algorithms: Naïve Bayes, Random Forest, and Support Vector Machines. The last of them achieved the most balanced results with the highest F1-score of 71.4%. The results might suggest that, provided with sufficient amount and quality of training data, even relatively simple bag-of-words-based method can be useful in figurative language identification task.

In Chapter 4, I described MIss RoBERTa WiLDe (Metaphor Identification using Masked Language Model with Wiktionary Lexical Definitions), a model designed for automatic metaphor detection. My method is logically consistent and supported theoretically, as utilizing literal basic meanings of words follows the guidelines of Metaphor Identification Procedure (MIP) and the concept of Selectional Preference Violation (SPV). I argue that there is no better source of purely literal word senses than the lexical definitions of said words. I have enhanced the existing algorithm [16] by introducing a different

kind of sentence representation and collecting dictionary definitions of the target words in a fully automatic manner. The results I achieved in the set of experiments suggest that implementing my ideas can lead to performance gains in metaphor identification task.

As indicated by Mao et al. [66], having access to large-scale textual resources using words only in their basic literal meanings could elevate the performance of the algorithms used for metaphor detection. However, metaphors are abundant in human language irrespective of the genre and, because of it, finding a perfect knowledge base of this kind does not seem possible. Nonetheless, I think that in comparison to other types of currently available resources, it is very likely that dictionaries are closest to that ideal.

Having witnessed that deep-learning-based language models combined with information conveyed by lexical definitions proved successful in token-level metaphor detection task in English, in future I plan to introduce a similar method for Japanese as well.

However, in my approach to figurative language identification in Japanese, the datasets I have been using so far are labeled on the sentence-level. If I wanted to use the same method as in the experiments performed on English data, I would have to obtain the ground truth labels corresponding to all the words comprised by each sentence in the datasets. In order to do so, I may employ a number of annotators, using one of the crowd-sourcing platforms available online.

Appendices

手 'hand, arm, paw', 足 'leg, foot', 頭・あたま 'head', 体・からだ 'body', 身 'body', 全身・ぜんしん 'whole body', 顔・かお 'face', 眉・まゆ 'eyebrow, brow', まゆげ 'eyebrow', 目眼 'eye', 耳みみ 'ear', 鼻はな 'nose', 口・くち 'mouth, maw', 唇・くちびる 'lips', 歯 'tooth', 舌・した 'tongue', 首・くび 'neck' (…)

Appendix 1. (Sample) Body Parts (BP).

よう, みたい, っぽい, ごとく・如く, ごとし・如し, ごとき・如き, 位・くらい・ぐらい, まるで, さぞ, いかにも・如何にも, あたかも・恰も・ 宛も, ほど, 宛ら・さながら (…)

Appendix 2. (Sample) Simile Indicators (SI).

中 'interior, inside, middle', うえ 'upper side, top, upper part, surface', 下・もと 'bottom, bottom part, under', 間・あいだ 'between', 内・うち 'inside', 外・そと 'outside', 前・まえ 'front, frontal part, in front of', 奥・おく 'the inner part, the depths', 後ろ 'the back, the rear' (…)

Appendix 3. (Sample) Spatial Position Indicators (PI).

色 'color', 藍, 青・あお 'blue, green', 青竹・あおたけ 'bluish green, malachite green', 赤・あか 'red, crimson, scarlet', 茜色, あかね 'madder', 緋色・あけいろ・ひいろ 'scarlet, cardinal, crimson', 浅葱 色・あさぎいろ 'pale greenish blue', 小豆色・あずきいろ 'reddish brown, russet', 亜麻色・あまい ろ 'ecru, beige', 鶯色・うぐいすいろ 'brownish green, olive green', 鶯茶・うぐいすちゃ 'greenish brown', 鬱金色・うこんいろ 'bright yellow', 江戸紫・えどむらさき 'royal purple, bluish purple', 葡萄色・えびいろ 'reddish brown, maroon', 臙脂色・えんじいろ 'dark red, crimson', カーキ色 'khaki', 貝紫色・かいむらさきいろ 'Tyrian purple, royal purple', 柿色・かきいろ 'persimmon color, dark orange', 韓紅・唐紅・からくれない 'crimson', 黄・きいろ 'yellow' (...)

Appendix 4. (Sample) Colors (CO).

安心・あんしん 'peace of mind, relief, sense of security, sense of safety', 不安・ふあん 'uneasiness, uncertainty, anxiety', 感謝・かんしゃ 'gratitude, gratefulness, appreciation', 驚愕・きょうがく 'astonishment, amazement, shock, fright', 興奮・こうふん 'excitement, arousal', 驚く・おどろく 'be surprised, be astonished, be amazed', 好奇心・こうきしん 'curiosity, inquisitiveness', 冷静・れいせい 'calm, cool, self-possessed', 焦る・あせる 'be hasty, act hastily, make (undue) haste', 不思議,・ふしぎ 'wonderful, wondrous, amazing, marvelous. strange', リラックス 'relax', 緊張・きんちょう 'tension, be tensed, nervous', 喜ぶ・悦ぶ・慶ぶ・歓ぶ・よろこぶ 'be glad, be happy, be pleased', 嬉しい・うれしい 'glad, joyful, delightful, happy', 幸せ・しあわせ 'happiness, happy, lucky, fortunate', 悲しい・かなしい 'sad, unhappy, sorrowful, depressed, depressing, tragic', 寂しい・淋しい・ さみしい・さびしい 'lonely, lonesome, isolated', 怒り・いかり 'anger, rage, fury, wrath', 怒る・おこる 'be angry, be furious, be mad', 感動・かんどう 'deep emotion, strong impression, sensation' (...)

Appendix 5. (Sample) Emotions (EM).

声・こえ 'voice', 水・みず 'water', 音・おと 'sound, noise', 光・ひかり 'light, beam, ray', 空・そら 'sky, air', 風・かぜ 'wind, breeze', 花・はな 'flower, blossom', 波・なみ 'wave', 火 'fire, flame, blaze', 海・うみ 'sea', 底 'bottom, bed (of a river)', 女・おんな 'woman, female, lady', 言葉・ことば 'word, expression, language', 雨・あめ 'rain', 雲・くも 'cloud', 肌・はだ 'skin', 美しい・うつくしい 'beautiful, pretty, attractive, lovely, charming, picturesque, sweet', 葉・はっぱ 'leaf', 冷たい・つめたい 'cold, cool, icy, coldhearded, indifferent', 雪・ゆき 'snow', 影・かげ 'shade, shadow' (...)

Appendix 6. (Sample) Frequently Figuratively Used (FF).

燃える・もえる 'burn, blaze, be in flames', 齧る・かじる 'bite, gnaw, nibble', あがる 'go up, rise, ascend', あく 'open, be opened', あしらう 'treat, handle', あてがう 'apply, fit', あてはまる 'apply, be applicable, be valid, fit', あてる 'put, place, apply', あなどる 'despise, disdain, scorn, look down on', あびる 'bathe in (sth), pour (sth) on oneself', あふれでる 'overflow', あふれる 'oveflow, spill over', いきれる 'get angry', いざなう 'invite', いじる 'play with, touch, tamper with', いらっしゃる 'come, visit, be', うっちゃる 'throw away, discard, abandon', うろたえる 'be flustered, be confused, be upset', おえる 'finish, complete, bring to an end', おきる 'get up, get out of bed', おののく 'shudder, tremble, shiver' (...)

Appendix 7. (Sample) Wago Verbs (WV).

Bibliography

[1] Abrams, M. H. and Harpham, G. *A glossary of literary terms*. Cengage Learning, 2014.

[2] Aozora Bunko digital library. https://www.aozora.gr.jp.

[3] Babieno, M., Rzepka, R., and Araki, K. "Evaluating Classification Algorithms for Recognizing Figurative Expressions in Japanese Literary Texts". In: *International Conference of the Pacific Association for Computational Linguistics*. Springer. 2019, pp. 181–188. DOI: https://doi.org/10.1007/978-981-15-6168-9_16.

[4] Babieno, M., Rzepka, R., Araki, K., and Dybala, P. "Comparing Conceptual Metaphor Theory-Related Features Using Classification Algorithms in Searching for Expressions Used Figuratively Within Japanese Texts". In: *International Joint Conference on Artificial Intelligence*. Springer. 2019, pp. 195–212. DOI: 10.1007/978-3-030-56150-5_10.

[5] Babieno, M., Takeshita, M., Radisavljevic, D., Rzepka, R., and Araki, K. "MIss RoBERTa WiLDe: Metaphor Identification Using Masked Language Model with Wiktionary Lexical Definitions". In: *Applied Sciences* 12.4 (2022). DOI: 10.3390/app12042081.

[6] Babieno, M., Takishita, S., Rzepka, R., and Araki, K. "Retrieving Metaphorical Sentences from Japanese Literature Using Standard Text Classification Methods". In: *Proceedings of the 60th Language Sense Engineering SIG conference*. 2018, pp. 51–59.

[7] Barcelona, A. "Clarifying and applying the notions of metaphor and metonymy within cognitive linguistics: An update". In: *Metaphor and Metonymy in Comparison and Contrast*. Ed. by R. Dirven and

R. Pörings. Mouton de Gruyter, 2003. DOI: 10.1515/9783110219197.2.207.

[8] Birke, J. and Sarkar, A. "A Clustering Approach for Nearly Unsupervised Recognition of Nonliteral Language". In: *11th Conference of the European Chapter of the Association for Computational Linguistics*. Trento, Italy: Association for Computational Linguistics, 2006, pp. 329–336.

[9] Black, M. "Metaphor". In: *Proceedings of the Aristotelian Society* 55 (1954), pp. 273–294.

[10] Boers, F. *Spatial prepositions and metaphor: A cognitive semantic journey along the up-down and the front-back dimensions*. Tübingen, Germany: Gunter Narr Verlag, 1996.

[11] Britannica Concise Encyclopedia. Chicago, IL, US / London, England / New Delhi, India / Paris, France / Seoul, South Korea / Sydney, Australia / Taipei, Taiwan / Tokyo, Japan: Encyclopaedia Britannica, Inc., 2008.

[12] Brown, T. et al. "Language models are few-shot learners". In: *Advances in neural information processing systems* 33 (2020), pp. 1877–1901.

[13] Caballero, R. and Ibarretxe-Antuñano, I. "Ways of perceiving, moving, and thinking: Revindicating culture in conceptual metaphor research". In: *Cognitive Semiotics* 5.1-2 (2009), pp. 268–290. DOI: 10.1515/cogsem.2013.5.12.268.

[14] Charniak, E. *BLLIP 1987-1989 WSJ corpus*. Philadelphia, Pa, 2000. DOI: 10.35111/fwew-da58.

[15] Chintalapudi, N., Battineni, G., and Amenta, F. "Sentimental Analysis of COVID-19 Tweets Using Deep Learning Models". In: *Infectious Disease Reports* 13.2 (2021), pp. 329–339. DOI: 10.3390/idr13020032.

[16] Choi, M., Lee, S., Choi, E., Park, H., Lee, J., Lee, D., and Lee, J. "MelBERT: Metaphor Detection via Contextualized Late Interaction using Metaphorical Identification Theories". In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online: Association for Computational Linguistics, 2021, pp. 1763–1773. DOI: 10.18653/v1/2021.naacl-main.141.

[17] Cornelissen, J. P. "Metaphor as a method in the domain of marketing". In: *Psychology and Marketing* 20.3 (2003), pp. 209–225. DOI: 10.1002/mar.10068.

[18] Corver, N. and Riemsdijk, H. van. "Semi-lexical categories". In: *Semi-lexical Categories: The Function of Content Words and the Content of Function Words*. Berlin, Germany / Boston, MA, US: De Gruyter Mouton, 2013, pp. 1–20. DOI: 10.1515/9783110874006.1.

[19] Davidson, D. "What metaphors mean". In: *Critical inquiry* 5.1 (1978), pp. 31–47. DOI: 10.1086/447971.

[20] Delahunty, G. P. and Garvey, J. J. *The English language: From sound to sense*. Anderson, SC, US: Parlor Press LLC, 2010.

[21] Devlin, J., Chang, M., Lee, K., and Toutanova, K. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*. Ed. by J. Burstein, C. Doran, and T. Solorio. Association for Computational Linguistics, 2019, pp. 4171–4186. DOI: 10.18653/v1/n19-1423.

[22] Dobrovol'skij, D. and Piirainen, E. *Figurative language: Cross-cultural and cross-linguistic perspectives*. Elsevier, 2005.

[23] Dobrzyńska, T. "Translating metaphor: Problems of meaning". In: *Journal of pragmatics* 24.6 (1995), pp. 595–604. DOI: 10.1016/0378-2166(95)00022-k.

[24] Fainsilber, L. and Ortony, A. "Metaphorical uses of language in the expression of emotions". In: *Metaphor and Symbolic Activity* 2.4 (1987), pp. 239–250. DOI: 10.1207/s15327868ms0204_2.

[25] Fass, D. "met*: A Method for Discriminating Metonymy and Metaphor by Computer". In: *Computational Linguistics* 17.1 (1991), pp. 49–90.

[26] Fass, D. and Wilks, Y. "Preference semantics, ill-formedness, and metaphor". In: *Computational Linguistics* 9 (June 2002), pp. 178–187.

[27] Fauconnier, G. and Turner, M. "Rethinking metaphor". In: *The Cambridge Handbook of Metaphor and Thought*. Ed. by R. W. J. Gibbs. Cambridge University Press, 2008, pp. 53–66. DOI: 10.1017/cbo9780511816802.005.

[28] Fellbaum, C. *WordNet: An Electronic Lexical Database*. Cambridge, MA, US: Bradford Books, 1998.

[29] Fogelin, R. J. "Metaphors, Similes and Similarity". In: *Aspects of Metaphor*. Springer Netherlands, 1994, pp. 23–39. DOI: 10.1007/978-94-015-8315-2_2.

[30] Gao, G., Choi, E., Choi, Y., and Zettlemoyer, L. "Neural Metaphor Detection in Context". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, 2018, pp. 607–613. DOI: 10.18653/v1/D18-1060.

[31] Gibbs Jr., R. W. "Process and products in making sense of tropes". In: *Metaphor and Thought*. Ed. by A. Ortony. 2nd ed. Cambridge University Press, 1993, pp. 252-276. DOI: 10.1017/CBO9781139173865.014.

[32] Gibbs Jr., R. W. *The Cambridge handbook of metaphor and thought*. Cambridge University Press, 2008. DOI: 10.1017/CBO9780511816802.

[33] Gibbs Jr., R. W. and Colston, H. L. "Chapter 7 Image schema". In: *Cognitive Linguistics: Basic Readings*. Ed. by D. Geeraerts. Berlin, Germany / New York, NY, US: De Gruyter Mouton, 2008, pp. 239–268. DOI: 10.1515/9783110199901.239.

[34] Goossens, L. "Metaphtonymy: The interaction of metaphor and metonymy in expressions for linguistic action". In: *Metaphor and Metonymy in Comparison and Contrast*. Walter de Gruyter, Berlin/New York, 1990. DOI: 10.1515/9783110219197.349.

[35] Green, E. D. "What are the most-cited publications in the social sciences (according to Google Scholar)?" In: *Impact of Social Sciences Blog* (2016).

[36] Group, P. "MIP: A Method for Identifying Metaphorically Used Words in Discourse". In: *Metaphor and Symbol* 22.1 (2007), pp. 1–39. DOI: 10.1080/10926480709336752.

[37] Haagsma, H. and Bjerva, J. "Detecting novel metaphor using selectional preference information". In: *Proceedings of the Fourth Workshop on Metaphor in NLP*. San Diego, California: Association for Computational Linguistics, 2016, pp. 10–17. DOI: 10.18653/v1/w16-1102.

[38] Hashimoto, C. and Kawahara, D. "Compilation of an idiom example database for supervised idiom identification". In: *Language resources and evaluation* 43.4 (2009), pp. 355–384. DOI: 10.1007/s10579-009-9104-1.

[39] Hashimoto, C., Sato, S., and Utsuro, T. "Japanese idiom recognition: Drawing a line between literal and idiomatic meanings". In: *Proceedings of the COLING/ACL on Main conference poster sessions*. Association for Computational Linguistics. 2006, pp. 353–360. DOI: 10.3115/1273073.1273119.

[40] Haspelmath, M. "Word Classes and Parts of Speech". In: *International Encyclopedia of the Social & Behavioral Sciences*. Elsevier, 2001, pp. 16538–16545. DOI: 10.1016/b0-08-043076-7/02959-4.

[41] Haspelmath, M. *From space to time: Temporal adverbials in the world's languages*. Munich: Lincom Europa, 1997. DOI: 10.5281/zenodo.831421.

[42] Hellsten, I. and Renvall, M. "Inside or outside of politics?: Metaphor and paradox in journalism". In: *Nordicom Review* 18.2 (1997), pp. 41–47.

[43] Heywood, J., Semino, E., and Short, M. "Linguistic metaphor identification in two extracts from novels". In: *Language and Literature: International Journal of Stylistics* 11.1 (2002), pp. 35–54. DOI: 10.1177/096394700201100104.

[44] Hiraga, M. K. "Tao of learning: Metaphors Japanese students live by". In: *Metaphors for Learning*. John Benjamins Publishing Company, 2008, pp. 55–72. DOI: 10.1075/hcp.22.06hir.

[45] Japanese Local Assembly Minutes Corpus Project. http://local-politics.jp.

[46] Japanese Wikipedia dumps list. https://dumps.wikimedia.org/jawiki/latest.

[47] JUMAN (Morphological Analyzer for Japanese). http://nlp.ist.i.kyoto-u.ac.jp/EN/index.php?JUMAN.

[48] Kageyama, T. and Kishimoto, H., eds. *Handbook of Japanese Lexicon and Word Formation*. De Gruyter Mouton, 2016. DOI: doi:10.1515/9781614512097.

[49] Kasanuki, Y. "Metaphor". In: *Cognitive Linguistics: From the Basics to the Forefront (in Japanese)*. Tokyo, Japan: Kuroshio Shuppan, 2013, pp. 53–78.

[50] Komatsubara, T. "The Corpus of Japanese Figurative Language: Toward". In: *Journal of Intercultural Studies (Kobe University)* 55 (2021).

[51] Kopp, R. R. *Metaphor therapy: Using client generated metaphors in psychotherapy*. New York, NY, USA: Routledge, 2013.

[52] Lakoff, G. "Metaphor and War: The Metaphor System Used to Justify War in the Gulf". In: *Cognitive Semiotics* 4.2 (2012), pp. 5–19. DOI: 10.1515/cogsem.2012.4.2.5.

[53] Lakoff, G. "Metaphor, morality, and politics, or, why conservatives have left liberals in the dust". In: *Social Research* (1995), pp. 177–213.

[54] Lakoff, G. "The Meanings of Literal". In: *Metaphor and Symbolic Activity* 1.4 (1986), pp. 291–296. DOI: 10.1207/s15327868ms0104_3.

[55] Lakoff, G. and Johnson, M. "Conceptual metaphor in everyday language". In: *The journal of Philosophy* 77.8 (1980), pp. 453–486. DOI: 10.2307/2025464.

[56] Lakoff, G. and Johnson, M. *Metaphors We Live By*. University of Chicago Press, 1980.

[57] Lakoff, G. and Johnson, M. "The Metaphorical Structure of the Human Conceptual System". In: *Cognitive Science* 4.2 (1980), pp. 195–208. DOI: 10.1207/s15516709cog0402_4.

[58] Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., and Soricut, R. "ALBERT: A Lite BERT for Self-supervised Learning of Language Representations". In: *International Conference of Learning Representations*. 2020.

[59] Leezenberg, M. "From cognitive linguistics to social science: Thirty years after Metaphors We Live By". In: *Cognitive Semiotics* 5.1-2 (2009), pp. 140–152. DOI: 10.1515/cogsem.2013.5.12.140.

[60] Leong, C. W. B., Beigman Klebanov, B., Hamill, C., Stemle, E., Ubale, R., and Chen, X. "A Report on the 2020 VUA and TOEFL Metaphor Detection Shared Task". In: *Proceedings of the Second Workshop on Figurative Language Processing*. Online: Association

for Computational Linguistics, 2020, pp. 18–29. DOI: 10.18653/v1/2020.figlang-1.3.

[61] Leong, C. W. B., Klebanov, B. B., and Shutova, E. "A Report on the 2018 VUA Metaphor Detection Shared Task". In: *Proceedings of the Workshop on Figurative Language Processing*. Association for Computational Linguistics, 2018, pp. 56–66. DOI: 10.18653/v1/w18-0907.

[62] Liu, Y. et al. "RoBERTa: A Robustly Optimized BERT Pretraining Approach". In: *CoRR* abs/1907.11692 (2019).

[63] Livedoor News. http://news.livedoor.com.

[64] Lu, X. and Wang, B. P.-Y. "Towards a metaphor-annotated corpus of Mandarin Chinese". In: *Language Resources and Evaluation* 51.3 (2017), pp. 663–694. DOI: 10.1007/s10579-017-9392-9.

[65] Maalej, Z. A. and Yu, N. *Embodiment via body parts: Studies from various languages and cultures*. Vol. 31. John Benjamins Publishing, 2011. DOI: https://doi.org/10.1075/hcp.31.

[66] Mao, R., Lin, C., and Guerin, F. "End-to-End Sequential Metaphor Identification Inspired by Linguistic Theories". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, 2019, pp. 3888–3898. DOI: 10.18653/v1/P19-1378.

[67] Mason, Z. J. "CorMet: A Computational, Corpus-Based Conventional Metaphor Extraction System". In: *Computational Linguistics* 30.1 (2004), pp. 23–44. DOI: 10.1162/089120104773633376.

[68] Masui, F., Fukumoto, J., Shiino, T., and Kawai, A. "A Method of Metaphoricity Detecion Using Probabilistic Measurements (in Japanese)". In: *Journal of Natural Language Processing* 9.5 (2002), pp. 72–92.

[69] Meier, B. P. "Do metaphors color our perception of social life?" In: *Handbook of Color Psychology*. Ed. by A. J. Elliot, M. D. Fairchild, and A. Franklin. Cambridge Handbooks in Psychology. Cambridge University Press, 2015, pp. 419-432. DOI: 10.1017/CBO9781107337930.021.

[70] Mhamdi, E. M. E., Guerraoui, R., and Rouault, S. "Distributed Momentum for Byzantine-resilient Stochastic Gradient Descent". In: *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

[71] Miller, E. F. "Metaphor and Political Knowledge". In: *American Political Science Review* 73.1 (1979), pp. 155–170. DOI: 10.2307/1954738.

[72] Mohammad, S., Shutova, E., and Turney, P. "Metaphor as a Medium for Emotion: An Empirical Study". In: *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*. Berlin, Germany: Association for Computational Linguistics, 2016, pp. 23–33. DOI: 10.18653/v1/S16-2003.

[73] Neidlein, A., Wiesenbach, P., and Markert, K. "An analysis of language models for metaphor recognition". In: *Proceedings of the 28th International Conference on Computational Linguistics*. Barcelona, Spain (Online): International Committee on Computational Linguistics, 2020, pp. 3722–3736. DOI: 10.18653/v1/2020.coling-main.332.

[74] Nekomoto, T., Sawayama, T., Nakajima, Y., Honma, H., Ptaszynski, M., Masui, F., and Masuyama, S. "Improving Accuracy of Metaphor Detection System with Heuristics (in Japanese)". In: *Proceedings of The 24th Annual Meeting of The Association for Natural Language Processing*. Okayama, Japan, pp. 89–92.

[75] Neuman, Y., Assaf, D., Cohen, Y., Last, M., Argamon, S., Howard, N., and Frieder, O. "Metaphor Identification in Large Texts Corpora". In: *PloS one* 8 (Apr. 2013), e62343. DOI: 10.1371/journal.pone.0062343.

[76] Onai, H. *Great Dictionary of 33800 Japanese Metaphors and Synonyms (in Japanese)*. Kodansha, 2005.

[77] Ortony, A. "The role of similarity in similes and metaphors". In: *Metaphor and Thought*. Cambridge University Press, 1993, pp. 342–356. DOI: 10.1017/cbo9781139173865.018.

[78] Ortony, A. and Fainsilber, L. "The role of metaphors in descriptions of emotions". In: *Proceedings of the 1987 workshop on Theoretical issues in natural language processing*. Association for Computational Linguistics, 1987. DOI: 10.3115/980304.980346.

[79] Palmatier, R. A. *Speaking of animals: A dictionary of animal metaphors*. Greenwood Publishing Group, 1995.

[80] Panther, K. U. and Radden, G. *Metonymy in language and thought*. Vol. 4. John Benjamins Publishing, 1999. DOI: https://doi.org/10.1075/hcp.4.

[81] Partington, A. "A corpus-based investigation into the use of metaphor in British business journalism". In: *Anglais de Specialité* 7-10 (1995), pp. 25–39. DOI: 10.4000/asp.3718.

[82] Pedregosa, F. et al. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[83] Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. "Deep Contextualized Word Representations". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, 2018, pp. 2227–2237. DOI: 10.18653/v1/N18-1202.

[84] Pollio, H., Barlow, J., Fine, H., and Pollio, M. *Psychology and the Poetics of Growth: Figurative Language in Psychology, Psychotherapy, and Education*. Mahwah, NJ, US: Lawrence Erlbaum Associates, 1977.

[85] Pütz, M. and Dirven, R., eds. *The Construal of Space in Language and Thought*. Berlin, Germany / New York, NY, US: De Gruyter Mouton, 2011. DOI: 10.1515/9783110821611.

[86] Radden, G. "The metaphor TIME AS SPACE across languages". In: *Zeitschrift für interkulturellen Fremdsprachenunterricht* 8.2 (2003), pp. 226–239.

[87] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. "Language models are unsupervised multitask learners". In: *OpenAI blog* 1.8 (2019), p. 9.

[88] Rai, S. and Chakraverty, S. "A Survey on Computational Metaphor Processing". In: *ACM Computing Surveys* 53.2 (2021), pp. 1–37. DOI: 10.1145/3373265.

[89] Reimers, N. and Gurevych, I. "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, Hong Kong, China, November 3-7, 2019*. Association for Computational Linguistics, 2019, pp. 3982–3992. DOI: 10.18653/v1/d19-1410.

[90] Schofer, P. and Rice, D. "Metaphor, Metonymy, and Synecdoche Revis(it)ed". In: *Semiotica* 21.1-2 (1977), pp. 121–150. DOI: 10.1515/semi.1977.21.1-2.121.

[91] Semino, E., Heywood, J., and Short, M. "Methodological problems in the analysis of metaphors in a corpus of conversations about cancer". In: *Journal of Pragmatics* 36 (July 2004), pp. 1271–1294. DOI: 10.1016/j.pragma.2003.10.013.

[92] Shutova, E., Sun, L., Gutiérrez, E. D., Lichtenstein, P., and Narayanan, S. "Multilingual Metaphor Processing: Experiments with Semi-Supervised and Unsupervised Learning". In: *Computational Linguistics* 43.1 (Apr. 2017), pp. 71–123. DOI: 10.1162/COLI_a_00275.

[93] Shutova, E., Sun, L., and Korhonen, A. "Metaphor identification using verb and noun clustering". In: *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*. 2010, pp. 1002–1010. DOI: 10.5555/1873781.1873894.

[94] Shutova, E., Teufel, S., and Korhonen, A. "Statistical Metaphor Processing". In: *Computational Linguistics* 39.2 (2013), pp. 301–353. DOI: 10.1162/coli_a_00124.

[95] Siegelman, E. Y. *Metaphor and meaning in psychotherapy*. New York, NY, USA: Guilford Press, 1993.

[96] Sommer, R. and Sommer, B. A. "Zoomorphy: Animal metaphors for human personality". In: *Anthrozoös* 24.3 (2011), pp. 237–248. DOI: 10.2752/175303711x13045914865024.

[97] Steen, G. *Finding Metaphor in Grammar and Usage*. John Benjamins Publishing Company, 2007. DOI: 10.1075/celcr.10.

[98] Steen, G., Dorst, A. G., Herrmann, J. B., Kaal, A. A., and Krennmayr, T. *VU Amsterdam Metaphor Corpus*. Oxford Text Archive. 2010.

[99] Steen, G., Dorst, L., Herrmann, J., Kaal, A., and Krennmayr, T. "Metaphor in usage". In: *Cognitive Linguistics* 21 (Nov. 2010), pp. 765-796. DOI: 10.1515/cogl.2010.024.

[100] Steen, G., Dorst, L., Herrmann, J., Kaal, A., Krennmayr, T., and Pasma, T. *A method for linguistic metaphor identification: From MIP to MIPVU*. Amsterdam, The Netherlands / Philadelphia, PA, US: John Benjamins Publishing Company, June 2010. DOI: 10.1075/celcr.14.

[101] Steen, G. "Metaphor: Stylistic approaches". English. In: *The encyclopedia of language and linguistics, Second Edition (Vol 8)*. Ed. by K. Brown. 8. Elsevier, 2006, pp. 51–57. DOI: 10.1016/b0-08-044854-2/00525-3.

[102] Sterkenburg, P. van, ed. *A Practical Guide to Lexicography*. Amsterdam, The Netherlands / Philadelphia, PA, US: John Benjamins Publishing Company, 2003. DOI: 10.1075/tlrp.6.

[103] Stevenson, A. *Oxford dictionary of English*. Oxford University Press, USA, 2010.

[104] Su, C., Fukumoto, F., Huang, X., Li, J., Wang, R., and Chen, Z. "DeepMet: A Reading Comprehension Paradigm for Token-level Metaphor Detection". In: *Proceedings of the Second Workshop on Figurative Language Processing*. Online: Association for Computational Linguistics, 2020, pp. 30–39. DOI: 10.18653/v1/2020.figlang-1.4.

[105] Turbayne, C. M. *The Myth of Metaphor*. Columbia, University of South Carolina Press, 1962.

[106] Uchiyama, K. and Ishizaki, S. "A disambiguation method for Japanese compound verbs". In: *Proceedings of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*. 2003, pp. 81–88. DOI: 10.3115/1119282.1119293.

[107] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. "Attention is All You Need". In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. NIPS'17. Long Beach, California, USA: Curran Associates Inc., 2017, pp. 6000-6010.

[108] Veale, T., Shutova, E., and Klebanov, B. B. "Metaphor: A computational perspective". In: *Synthesis Lectures on Human Language Technologies* 9.1 (2016), pp. 1–160. DOI: 10.2200/S00694ED1V01Y201601HLT031.

[109] Wan, H., Lin, J., Du, J., Shen, D., and Zhang, M. "Enhancing Metaphor Detection by Gloss-based Interpretations". In: *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. Association for Computational Linguistics, 2021. DOI: 10.18653/v1/2021.findings-acl.173.

[110] Warren, B. "An alternative account of the interpretation of referential metonymy and metaphor". In: *Metaphor and Metonymy in Comparison and Contrast*. Ed. by R. Dirven and R. Pörings. Mouton de Gruyter, 2003. DOI: 10.1515/9783110219197.1.113.

[111] Wilks, Y. "Making preferences more active". In: *Artificial Intelligence* 11 (3 1978), pp. 197–223. DOI: 10.1016/0004-3702(78)90001-2.

[112] Wilks, Y., Dalton, A., Allen, J., and Galescu, L. "Automatic Metaphor Detection using Large-Scale Lexical Resources and Conventional Metaphor Extraction". In: *Proceedings of the First Workshop on Metaphor in NLP*. Atlanta, Georgia, US: Association for Computational Linguistics, 2013, pp. 36–44.

[113] Yamanashi, M. *Metaphor and Understanding (in Japanese)*. Tokyo Daigaku Shuppan-kai, Tokyo, 1988.

[114] Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., and Le, Q. V. "XLNet: Generalized Autoregressive Pretraining for Language Understanding". In: *Advances in Neural Information Processing Systems* 32 (2019).

[115] Yu, N. "Spatial Metaphors for Morality: A Perspective from Chinese". In: *Metaphor and Symbol* 31.2 (2016), pp. 108–125. DOI: 10.1080/10926488.2016.1150763.

Acknowledgements

I would like to express my gratitude to:

Prof. Kenji Araki

for his unstinting and continued support throughout all my years in the Language Media Laboratory;

Assistant Prof. Rafał Rzepka

for his great patience and invaluable guidance on all my research projects and endeavors;

Members of the review committee:

Prof. Yuji Sakamoto, Prof. Miki Haseyama,

Prof. Yoshinori Dobashi, and Associate Prof. Toshihiko Ito

for their revealing insights and help in refining this dissertation;

Prof. Romuald Huszcza

for his infinite kindness, priceless wisdom and remarkable efforts in leading me to where I managed to arrive at;

Prof. Motoki Nomachi

for his truly generous help during my first years at Hokkaido University;

Associate Prof. Jarosław A. Pietrow

for his extraordinary support at University of Warsaw and far beyond;

Dušan Radisavljević and Masashi Takeshita

for their important assistance at the final stage of this research;

My Mom Grażyna and Dad Tadeusz

for much more than could ever be listed, especially for being my dearest friends, loyal and reliable even in the darkest hours;

My Wife Yuki

for loving, understanding, and fixing me (and being super cute ☺);

Saint Rita of Cascia

for firmly delivering my prayers to Heaven;

and

God Almighty

for giving me more than seventy-seven chances.