Figurative Language in Swedish Clinical Texts

Dimitrios Kokkinakis
Centre for Language Technology and the Swedish Language Bank,
University of Gothenburg, Sweden
dimitrios.kokkinakis@svenska.gu.se

Abstract

Automated processing of clinical texts is commonly faced with various less exposed, and not so regularly discussed linguistically complex problems that need to be addressed. One of these issues concerns the usage of figurative language. Figurative language implies the use of words that go beyond their ordinary meaning, a linguistically complex and challenging problem and also a problem that causes great difficulty for the field of natural language processing (NLP). The problem is equally prevalent in both general language and also in various sublanguages, such as clinical medicine. Therefore we believe that a comprehensive model of e.g. clinical language processing needs to account for figurative language usage, and this paper provides a description, and preliminary results towards this goal. Since the empirical, clinical data used in the study is limited in size, there is no formal distinction made between different sub-classifications of figurative language. e.g., metaphors, idioms or simile. We illustrate several types of figurative expressions in the clinical discourse and apply a rather quantitative and corpus-based level analysis. The main research questions that this paper asks are whether there are traces of figurative language (or at least a subset of such types) in patient-doctor and patient-nurse interactions, how can they be found in a convenient way and whether these are transferred in the electronic health records and to what degree.

1 Introduction

Automated processing of clinical texts with the intention to link all important text fragments to various established terminologies and ontologies for relation or event extraction is commonly faced with various less exposed, and not so regularly discussed linguistically motivated issues that needs to be addressed. One of these issues is the usage of figurative language. Figurative language, that is the use of words that go beyond their ordinary meaning, is not only a linguistically complex and challenging problem but also a problem that causes great difficulty for the field of natural language processing (NLP), both for the processing of general language and of various sublanguages, such as clinical medicine. Therefore we believe that a comprehensive model of e.g. clinical language processing needs to account for figurative language usage, and this paper provides a description, and preliminary results towards this goal. Since the empirical, clinical data used in the study is limited in size, there is no formal distinction made between different sub-classifications of figurative language. e.g., metaphors, idioms or simile. As a matter of fact, all these types of expressions form a continuum with fuzzy boundaries [9], and most of the NLP-oriented approaches discussed in the past have used either very large data for the analysis or hand annotates samples [17], a situation that has been prohibitive so far in our project. Therefore distinction is solely based on a more general level, namely between literal versus figurative language, and on a more quantitative and corpus-based level, supported with concrete examples that illustrate several types of figurative expressions in the clinical discourse. The main research questions that this paper asks are whether there are traces of figurative language (or at least a subset of such types) in patient-doctor and patient-nurse interactions, how can they be found in a convenient way and whether these are transferred in the electronic health records and to what degree.

2 Theoretical Background (Idioms, Metaphors and Similes)

Figurative language has been in a focus among several scholars during a very long period. Cognitive linguists, for instance ([12], [11]), have been studying figurative language for many years under the assumption that word meaning is not ”fixed” but is rather a function of perspective, therefore compositionality (the degree to which the features of the parts of a multi word expression combine to predict the features of the whole) can only be achieved if context is taken into consideration. Figurative language, compared to literal language, refers to simple words, and most frequently, to multiword expressions that deviate from their defined meaning by e.g., exaggerating or altering the usual meanings of the constituent words [18]. Figurative language, and its realization into various rhetorical devices, i.e. figures of speech, departs from literal meaning to achieve some form of a particular effect on the listener or reader.

There are several different, but related, types of such figures of speech but here we mainly investigate just three of the most common ones, namely idioms, metaphors and similes using a combination of simple and flexible automated techniques. An *idiom* (e.g., ”He is pulling my leg”) is an expression consisting of a combination of words that have a figurative meaning. Idiomatic expressions are usually presumed to be figures of speech contradicting the principle of compositionality. Many idioms appear in language first as metaphors and if during some time period, a large part of a population finds them interesting, and/or useful, then they become a fixed part of the language as idioms. A *metaphor* (e.g., ”All the world’s a stage”) is a literary figure of speech that describes a subject by asserting that it is, on some point of comparison, the same as another otherwise unrelated object. A metaphor asserts that one thing is another thing, not just that one is *like* another, such as in the case of *simile* (e.g., ”He fights like a lion”) which is a figure of speech that directly compares two different things. Similes usually employ the words ”as” and ”like” or other comparative words such as ”than”. A simile differs from a metaphor in that the latter compares two unlike things by saying that the one thing is the other thing. Similes can be negative, too, asserting that two things are unlike in one or more respects. Finally, previous work has shown that common figures of speech, such as idioms and metaphors, also involve some degree of lexical and syntactic variability, in the sense that, for instance, some allow pluralization while some not.

3 Previous Research

Some of the work that has been conducted from a strong corpus based and/or NLP perspective includes the work by [8] who focus on a particular class of English idiomatic expression, i.e., those that involve the combination of a verb plus a noun in its direct object position. Fazly & Stevenson [8] investigated a lexicon-based method were a lexical and syntactic flexibility (e.g., verbal inflection; pluralization; internal modification; use of determiner types; passivization) was allowed in a restrictive form. Among the idioms examined, some exhibited limited morphosyntactic flexibility, while others were more syntactically flexible. For instance, the idiom ”shoot the breeze” can undergo verbal inflection, i.e. ”shot the breeze”, but not internal modification or passivization, i.e. ”the breeze was shot” [4]. A main interpretative approach to non-literal use detection (e.g., metaphors) is the investigation of whether there are some sort of selectional preference violations in a given context; *cf.* [19] and [15]. According to Wilks, selectional restrictions are the semantic constraints that a verb places onto its arguments, ”a car drinks gasoline”, since the English verb ”to drink” tends to have a human or animal as the subject and food or potable liquid as the direction object, respectively. In a similar path, Hank [10], in his example ”political storm”, discusses how adjectives can be classified in order to pick out an appropriate subset, by first distinguishing between adjectives that identify kinds of storms as natural phenomena and those where the word is used metaphorically. If the correct interpretation of e.g. ”storm” is to be activated, it is necessary first to distinguish between adjectives that identify storms as natural phenomena and then those where the word is used in a non-literal sense. Other work is specifically geared towards metaphor recognition [14]. For instance, two kinds of metaphorical language is distinguished, ”conventional metaphors”, mostly idiomatic expressions, that become a part of an ordinary discourse, and ”creative metaphors”, which are distinctly novel or ad hoc uses of language, since neither the creator nor the audience has encountered

the metaphor before. Creative metaphors are frequently used in conversation to describe an emotional experience [5]. Birke & Sarkar [2] presented a system for automatically classifying literal and non literal usages of phrasal and expression verbs, for example ”throw away”, through nearly unsupervised word-sense disambiguation and various clustering techniques based on models build on hand-annotated data. Finally, figurative usage seems to be an important research topic in the medical domain and there is some evidence that supports this claim. For example, it has been shown that figurative language has an active role in the narratives of patients with cancer. Such non-literal usage, e.g. metaphors, can bridge the gap between the cancer experience and the world of technology and treatment, helping patients to symbolically control their illness [13] while other studies have looked at how non-literal language shifts throughout a person's recovery period [3]. We believe that detection of figurative language can play an important role both for the deep understanding of the discourse and communication interplay, and particularly, also for the part of NLP that involves automatic text understanding, where figurative language is considered a serious bottleneck [16].

4 Experimental Setting

This study is part of an ongoing larger project which investigates how complex, vague and long-standing symptoms with no identified organic cause are put into context, interpreted and acted upon in primary health-care interactions. It is based on studying interactions between patients and nurses giving advice over telephone, consultations between patients and physicians, interviews and study patients' medical records and case notes. Eighteen eligible patients who have contacted their primary health care centre by telephone, have had at least eight physical consultations with nurses or physicians in the last 12 months and with a majority of the symptoms within this time span with no clear organic or psychiatric cause were selected for the project. At the end, and due to some practical considerations only data from 16 patients was finally used (75,000 tokens). The overall expected results is to facilitate the development of future interventions aimed at decreasing the morbidity due to Medically Unexplained Symptoms (MUS) and give further insights into the problem. There is no generally accepted diagnostic criteria for such Medically Unexplained Symptoms, but one proposed definition is ”one or more physical symptoms which have been present at least three months, cause the patient clinically significant distress or impairment and cannot be explained by any recognizable physical disease” [7]. Common symptoms among MUS patients are: *headaches, dizziness, fatigue, dyspepsia, bloatedness, myalgia, joint pain, facial pain, pelvic pains, lower back pain, cervical pain and slowness of thoughts*.

The chosen method is corpus based and is geared onto a rather quantitative analysis of the figurative language phenomenon in Swedish clinical texts. The approach we follow is basically a heuristic pattern matching approach which looks at lexically derived idiomatic keywords and their order in a sentence, with certain variation allowed. If a string fragment corresponding to a particular rule or pattern is found it is annotated. All sentences with annotations are then manually analyzed. The main instruments we make use of are large lists of available idiomatic expressions (idioms and metaphors) extracted from various monolingual lexical resources for written Swedish, roughly 6,000 idiomatic expressions as well as from several Internet sites, e.g., from `http://www.livet.se/ord/k%C3%A4lla/Idiom/`. We make also use of a list of string matching patterns, essentially rules, manually developed particularly for similes. The extracted list of idiomatic expressions is modeled as a finite state recognizer which permits a controlled form of lexical and syntactic variability. This variability is modeled as regular expressions with optional slot fillers that may or may not be filled by string fragments during processing. For instance, for the Swedish idiom: *tappa sugen* (lit. ”drop the craving”, i.e., ”give up”, ”lose interest”) we allow both inflection for the verb *tappa* (e.g., *tappade*, i.e., ”lost”) and also the possibility of intervening other words, usually one to three arbitrary words (e.g., *tappa inte sugen*, i.e. ”not give up”; *tappa hon inte sugen*, i.e. ”she did not give up”). For similes we use a limited list of characteristic single words or very short combinations of words, and their part of speech, in these rules, for instance *är som en/ett*, i.e. ”is like a”; *som en/ett*, i.e. ”like a” and *likt* or *liksom*, i.e. ”like”. These are modeled in a similar manner as before, by using regular expression patterns. Such a pattern (in a simplified form) may look like the

following way: *Determiner? Adjective* (Pronoun/Noun) (är/other verbs) som (en/ett) Adjective* Noun* (where the symbol ”/” is meant here to function as a disjunction). Here, however, we do not allow much variability since there is a significant risk of allowing the recognition of a large number of false positives and spurious results. However, we do allow gender variability *en* or *ett* and limited inflection.

5 Results

The results from the application of the previously outlined method were manual analyzed in order to get deeper insights into: a) the performance of the automatic recognition (predominantly from a *precision score* perspective) of the approach b) get an idea of the magnitude of the identified figurative expressions in the clinical data and c) get some guidance on whether figurative expressions are transferred from the patient-doctor and patient-nurse interactions into the electronic health records and to what degree. Finally, we would also like to find out which new figurative expressions are in the data and not captured by the existing resources; for instance are these *creative metaphors* or other types. However, due to time constraints this topic was not prioritized for the time being and thus not elaborated in detail.

Although the available data is limited in size there was a fairly large number of instances that could be identified, and some of those were rather creative, interesting and relevant for their context. In the electronic health records 143 (69 diferent) figurative, multiword expression could be identified (10 similes). Examples of both include: *ta sitt liv*, i.e. lit: ”take your life”, ”commit suicide”; *kasta vatten*, i.e. lit. ”to throw water”, ”to urinate” and *gå ner i vikt*, i.e. lit. ”to go down in weight”, ”to lose weight”; and for similes *...ont som ett sug i magen* i.e. ”...hurts like a suction in the stomach”. In the transcribed interactions, which were much limited in size, 105 (42 diferent) figurative expression could be identified (21 similes). Examples of both include: *ett oskrivet blad*, i.e. lit. ”an unwritten paper”, ”pure and innocent”; *det spelar ingen roll*, i.e. lit. ”it does not play any role”, ”it does not matter”; and *hålla tummarna*, i.e. lit. ”to hold the thumbs”, ”to wish for luck”; and for similes *diskbråck ... det är som en blixt*, i.e. ”intervertebral disc displacement ... is like a flash from the sky”.

From the manual analysis of all annotations (248), we could obtain an overall precision score of 72.1% (calculated as *Precision = true positives/true positives+false positives*; recall was not measured). The number of false positives (items incorrectly labeled) was 69 and the majority were falsely annotated similes triggered by the designed patterns, since due to their rather very general nature also identified a number of spurious candidates; for instance *...i Aquacel som en tamponad*, i.e. ”...adds the Aquacel like/as a tamponade” or *...det är en rädsla*, i.e. ”...it is a fear” (unclear pronominal reference). Finally, there was a 10.8% overlap between the annotation in the records and the transcribed dialogues.

6 Discussion and Future Work

There are several different types of figurative language in various discourse contexts with a high frequency of use, and in this paper we have investigated whether a subset of such expressions exist in Swedish clinical data. The preliminary results showed some very clear traces of such language and a large number could be identified using available lexical resources with high precision. However, recall is also important and future plans also include means to identify novel idiomatic expressions and/or other types of figurative language that seem to prevail in some of the clinical data. For instance, in the transcribed dialogues there is an overwhelming number of onomatopoetic expressions (i.e., imitation of a sound made by or associated with its referent, such as ”wow” and ”pff”). We would like to find out which new figurative expressions are in the data and not captured by the existing resources. Are these for instance *creative metaphors* or other types and of what kind. Through preliminary manual analysis we could identify several novel figurative expressions and predominantly in the transcribed dialogues, such as: *emotionell rutschelkana*, i.e. ”emotional slide”; *går genom svarta hål*, i.e. ”go through black holes”; *lärt känna nya sidor av dig själv*, i.e. ”get to know new sides of yourself”; *sakta bygga från fötterna till huvud*, i.e. ”slowly build from the feet to the head”; *bakom den människa/person*, i.e. ”behind that

person" and a smaller number in the records, e.g. *väldigt utbränd*, i.e. "very burned out". In such interactions, the ability to recognize and understand patients use of figurative language can provide clinicians with means of evaluating personality and styles of thinking.

As a future task it would be also highly relevant to qualitatively investigate the reasons why figurative language is present in the health records. Does it depend on lack of understanding from the coders side (usually a contact nurse); is it simply a convenient way to describe symptoms and states; is there lack of appropriate nomenclature? Moreover, an idiom type often has a literal interpretation as well. Therefore, the exploration of e.g. use of informative prior knowledge about the overall syntactic behavior of potentially-idiomatic expressions to determine whether an instance of the expression is used idiomatically or not, is of great importance for many (semantically oriented) NLP applications [6], an issue that requires more studies, particularly in critical domains where the distinction can have severe consequences. Moreover, the identification of figurative expressions, which describe physical or emotional symptoms, is a very useful supporting component, since these important expressions can be then automatically linked to existing medical ontologies and enhance e.g., decision support or other systems.

Currently, the experimentation is based on limited amount of data, therefore it is difficult to draw clear conclusions as to the magnitude of the impact the ability to identify idiomatic and figurative expressions would have on improving medical NLP or clinical care delivery. However, larger scale studies for other languages and domains have shown to be useful in many applications. Moreover, like sentiment analysis or opinion extraction, computational figurative identification can provide an understanding of the framings or conceptualizations used in various communities or subdomains [1].

References

[1] E. Baumer and B. Tomlinson. Computational metaphor identification in communities of blogs. *ICWSM. Proceedings of the Second International Conf. on Weblogs and Social Media. Seattle, USA, AAAI Press.*, 2008.

[2] J. Birke and A. Sarkar. A clustering approach for the nearly unsupervised recognition of non-literal language. *Proceedings of the 11th EACL. Trento, Italy.*, 30:329–336, 2006.

[3] C. Boylstein, M. Rittman, and R. Hinojosa. Metaphor shifts in stroke recovery. *Health Commun.*, 21(3):279–87, 2007.

[4] LJ. Brinton and E. Closs Traugott. Lexicalization and language change. *Cambridge University Press.*, page 55, 2005.

[5] R. Carter. *Language and creativity: The art of common talk.* Routledge, New York., 2004.

[6] P. Cook, A. Fazly, and S. Stevenson. Pulling their weight: Exploiting syntactic forms for the automatic identif ication of idiomatic expressions in context. *Proceedings of the ACL Workshop on A Broader Perspective on Multiword Expressions. Prague.*, pages 41–48, 2007.

[7] R. Peveler R. *et al.* Medically unexplained physical symptoms in primary care: a comparison of self-report screening questionnaires and clinical opinion. *J Psychosom Res.*, 42(3):245–252, 1997.

[8] P. A. Fazly and S. Stevenson. Automatically constructing a lexicon of verb phrase idiomatic combinations. *Proceedings of the 11th EACL. Trento, Italy.*, pages 337–344, 2006.

[9] RW. Gibbs. Literal meaning and psychological theory. *Cognitive Science*, 8:275–304, 1984.

[10] P. Hanks. The syntagmatics of metaphor and idiom. *International J of Lexicography*, 17:3, 2004.

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

[12] R. W. Langacker. *Concept, Image, and Symbol. The Cognitive Basis of Grammar.* Berlin: Mouton de Gruyter., 1990.

[13] C. Laranjeira. The role of narrative and metaphor in the cancer life story: a theoretical analysis. *Med Health Care Philos.*, 2012.

[14] Group Pragglejaz. Mip: A method for identifying metaphorically used words in discourse metaphor and symbol. *Lawrence Erlbaum Associates, Inc*, 22(1):1–39, 2007.

[15] P. Resnik. Selectional constraints: an information-theoretic model and its computational realization. *Cognition*, 61:127–159, 1996.

[16] E. Shutova. Models of metaphor in nlp. *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. Uppsala, Sweden.*, pages 688–697, 2010.

[17] E. Shutova and S. Teufel. Metaphor corpus annotated for source - target domain mappings. *Proceedings of the Conference on Language Resources and Evaluation (LREC). Malta.*, pages 3255–3261, 2010.

[18] L. Sikos, S. Windisch Brown, A. E. Kim1, L. A. Michaelis, and M. Palmer. Figurative language: 'meaning' is often more than just a sum of the parts. *Association for the Advancement of AI (AAAI). Conf. on Biologically Inspired Cognitive Architectures (BICA). Virginia, USA.*, 2008.

[19] Y. Wilks. Making preferences more active. *Artificial Intelligence. Reprinted in N. V. Findler, Associative Networks. New York: Academic Press.*, 11(3), 1978.