

Identification and Contextual Semantic Retrieval of Polysemy Words

Shashank U, Venkatesh B N, Rajeshwari S B, Jagadish S Kallimani

ABSTRACT--- Polysemy is a spectacle where an individual word is associated with two or more different meanings. The distinction between polysemy and monosemy is that in monosemy, there is no presence of semantic ambiguity in language that is a word will have only one meaning. Polysemy escalates in natural language: pragmatically every word is polysemy up to a certain extent. This paper attempts to disambiguate a polysemy word and classify it to its correct sense based on its use in particular context. By taking into account the words which surround the polysemy word, the article tries to get the right sense out of it. Oxford dictionary is used to fetch the meaning of polysemy word. In this article, we use the information rich phrase or sentence as a key to identify polysemy word and illuminate it to its appropriate sense. The article is a novel approach to identify a polysemy word just like how hearers effortlessly reach at the contextually appropriate sense of word on a given occasion of use.

Index Terms — Polysemy, Semantic Ambiguity, Neighboring Words, Contextual Sense, Oxford Dictionary

1. INTRODUCTION

Polysemy is identified as the phenomenon where a single word form is associated with two or more related senses. Polysemy is the ability of words, symbols and signs to have multiple meanings, and polysemy is a word or phrase that has multiple etymologically related meanings. Homonymous words come into picture when polysemy words are being talked about. How homonymous words differ from polysemy words is that, homonymous word is the one with unrelated meanings and polysemy word is the one with different related meanings. Thus the difference between homonyms and polysemy is subtle. Polysemy is contrasted with monosemy as well. Monosemy word has only one meaning. Although the distinctions between polysemy, monosemy and homonymy may seem clear at an intuitive level, they have proven difficult to draw in practice.

Polysemy is usually characterized as the phenomenon whereby a single word form is associated with two or several related senses, as in (1) below:

- (1) Draw a line; read a line; a line around eyes; a wash on a line; wait in a line; a line of bad decisions, etc.

In this, it is contrasted with monosemy, on the one hand, and with homonymy, on the other. While a monosemy form has only one meaning, a homonymous form is associated with two or several unrelated meanings (e.g., coach; 'bus',

'sports instructor'), and is viewed as involving different lexemes (e.g., COACH1, COACH2).

Polysemy is essential in natural languages, and it affects both content and function words. While deciding which sense is intended on a given occasion of use rarely seems to cause any difficulty for speakers of a language, polysemy has proved notoriously difficult to treat both theoretically and empirically.

The definition and delimitation of the polysemy phenomenon itself also remains a source of theoretical discussion across disciplines: how do we tell polysemy apart from monosemy on the one hand, and from homonymy on the other? At first glance, the contrast with monosemy is clearer: while a monosemy term has only a single meaning, a polysemy term is associated with several senses.

Polysemy proliferates in natural language: virtually every word is polysemy to some extent. Still, the phenomenon has been largely ignored in the mainstream linguistics literature, and in related disciplines such as philosophy of language. However, polysemy is a topic of relevance to linguistic and philosophical debates regarding lexical meaning representation, compositional semantics, and the semantics-pragmatics divide.

Judgements of polysemy are difficult to make as it is a vague concept of relatedness. Etymology is not an infallible test for polysemy because some unrelated words share a common historical origin. Moreover the test depends on speaker's judgments on relatedness. One of the several tests for polysemy includes zeugma: if one word seems to exhibit zeugma when applied in different contexts, it is likely that the contexts bring out different polysemy of the same word. If the two senses of same words do not seem to fit, yet seem related, then it is likely that they are polysemy.

However, such tests for identity of meaning do not give clear-cut answers. In particular, only a slight manipulation of the context can yield a different result, as shown by the following example:

- (2) a. Judy's dissertation is thought provoking though yellowed with age.
b. Judy's dissertation is still thought provoking though yellowed with age.

While the sentence in (2a) is zeugmatic – apparently due to the use of Judy's dissertation to refer to a type of informational content in the first conjunct and a physical object in the second conjunct – no zeugmatic effect occurs when the sentence is slightly altered as in (2b). Furthermore, the tests typically do not distinguish between polysemy and

Revised Manuscript Received on July 10, 2019.

Shashank U, Department of Information Science and Engineering, M S Ramaiah Institute of Technology, Bangalore, India.

Venkatesh B N, Department of Information Science and Engineering, M S Ramaiah Institute of Technology, Bangalore, India.

Rajeshwari S B, Department of Information Science and Engineering, M S Ramaiah Institute of Technology, Bangalore, India.

Jagadish S Kallimani, Department of Computer Science and Engineering, M S Ramaiah Institute of Technology, Bangalore, India. (E-mail: jagadish.k@msrit.edu)

homonymy – that is, they do not distinguish between senses or meanings that are related and those that are unrelated – both of which come out as instances of a more general phenomenon of lexical ambiguity. The fact that this test again depends on speaker’s judgments on relatedness, however, means that this test for polysemy is not infallible, but is rather merely a helpful conceptual aid.

It is customary in the literature to distinguish between regular or logical polysemy, on the one hand, and irregular or accidental polysemy, on the other. In formal semantic and computational approaches, regular polysemy is typically analyzed as being generated by lexical rules, in this way accounting for the productivity and cross-linguistic availability of the patterns of sense extension and at the same time avoiding a listing of all senses for the words. Irregular polysemy, on the other hand, is described as cases where the semantic distinction between the meanings for a word cannot be found in any other word of the given language. The English verb ‘run’ may be an example of this: its different senses in run a mile, run a shop, run late, run on gasoline, etc. seem idiosyncratic to this particular lexical item, and may each have arisen as a result of different lexical semantic or pragmatic processes, such as for instance specification, loosening, metaphorical extension, and so on.

However, the distinction between regular and irregular polysemy is not clear-cut either. As to irregular polysemy, there appears to be degrees of irregularity, with some cases being clearly idiosyncratic, and others constrained by the way meaning chains tend to develop.

Identification of polysemy words is one of the important works in any language for its technological development. Several related works are carried out in this field. Various papers, consisting of methods to solve this problem, are presented by various scholars [1],[2] and [3].

The aim of this research was to develop a solution for identification of polysemy word. Similar works have been carried out mainly by [3] and [4] by using computational linguistic tools like Wordnet.

This paper provides a brute force approach to identify polysemy word and finding its appropriate sense based on its use in a context. The method keeps track of words that surround ambiguous word and accordingly the word’s meaning will be retrieved from Oxford dictionary, the vocabulary tool used in this approach. There will be many meanings for the ambiguous word available in the dictionary, but our methodology picks the right meaning for that ambiguous word by considering its neighboring words.

2. RELATED WORKS

Recent work on polysemy is as varied as is the phenomenon itself, both in its focus and methods. In general linguistics, polysemy received little attention for many years, mainly due to the predominance of generative grammar with its focus on the sentence as the central unit of meaning. However, with the emergence of the cognitive grammar during the 1980s polysemy emerged on the research agenda as a key topic in lexical semantics, in particular as a result of the pioneering studies conducted by George Lakoff (1987) and Claudia Brugman (1988) on the polysemy of English prepositions. Alongside the cognitive linguistic movement, polysemy has become a central topic

of investigation within many formal and computational semantic approaches, starting with Pustejovsky’s (1995) seminal work on the topic and most recently culminating in Asher’s (2011) monograph *Lexical Meaning in Context*. With their focus on semantic compositionality, these accounts have focused mainly on logical polysemy, which seems to be more tractable from a formal/computational point of view. In addition to these two main trends in the research on polysemy, much of the work conducted within the relatively new field of lexical pragmatics has a direct bearing on the topic (e.g., Carston, 2002; Recanati, 2004; Wilson & Carston, 2007). These approaches are mainly concerned with how polysemy relates to the interaction between linguistically-encoded content and contextual information in the derivation of speaker-intended meanings. In the psycholinguistic literature, polysemy has attracted interest due to the issues it raises for semantic representation, in particular, how the mental lexicon represents polysemy compared with homonymy, a distinction that has been investigated using different methods and techniques.

An approach to identify the most frequent sense of a polysemy word is proposed in [1]. They have introduced two concepts that can benefit detection: companions, which are most frequently co-occurring words, and the most frequent translation in a bi-text. An approach for word sense disambiguation is proposed in [2]. In this approach, by using Wordnet, domains of each single word will be defined and a process of defining the best domain to be assigned to that particular word will be carried out. An approach for word sense disambiguation is also proposed in [3]. They have presented a knowledge based algorithm for disambiguating polysemy words using computational linguistics tool, Word Net. The task of word sense disambiguation requires finding out the similarity between the target word (word to be disambiguated) and the nearby words (words surrounding the target word in input text). An approach for identifying polysemy word is discussed in [4]. They propose a concept of context-related semantic set to identify the meaning of a word by considering the relations between the word and its contexts. Copestake and Briscoe (1995) suggest that the universal grinder (Pelletier, 1975), as well as a several conventionalized [5] sub-cases of it (meat-grinding, furring, and so on) might apply in typical instances of regular polysemy such as the following:

- (3) a. There was rabbit all over the highway. (Universal grinding)
- b. Steven had rabbit for dinner. (Meat-grinding)
- c. The model wore rabbit on the catwalk. (Fur-grinding)

In the examples in (3), the effect of the rules would be to create from a count noun denoting a physical object a mass noun with properties appropriate for not individuated substance (e.g., meat, fur, or general ‘stuff’). The rules are seen as coming with specific interpretive predictions based on lexically stored information, so that, for instance, a mass use of an animal term would have a ‘meat’ sense as default.



Evans distinguishes between three types of polysemy [6]: conceptual polysemy, lexical polysemy, and inter-lexical polysemy. These are illustrated below:

- (4) That book is heavy/illegible/boring/long. ('tome'/'text'/'level of interest'/'duration').
- (5) We are in a room/in pain. ('container'/'state')
- (6) a. We are in pain/in a room ('state'/'spatial')
- b. We are on the run/on the sand. ('state'/'spatial').

According to Evans, in (4) we have an instance of conceptual polysemy, in which an open-class lexical item (book) takes on slightly different interpretations in different contexts. The lexical polysemy in (5) involves, according to Evans, distinct lexical concepts conventionally associated with the preposition 'in'.

Finally, the novel concept of inter-lexical polysemy, exemplified by (6), involves systematic similarities between distinct lexical concepts associated with distinct lexical forms.

3. PROPOSED POLYSEMY ALGORITHM

The algorithm involves the following steps to disambiguate the sense of polysemy word:

- Accept a sentence containing polysemy word
- Identify the polysemy word in a sentence
- Remove stop-words in the sentence and perform stemming of polysemy word
- Retrieve all possible meanings of the identified polysemy word using oxford dictionary
- Providing the contextual meaning of polysemy word by using the words which surround the polysemy word as clue words

4. THE PROPOSED ARCHITECTURE

- *Context Module*: This module consists of input text which contains the polysemy word.
- *Tokenizer*: For our language processing, we want to break up the input text into words and punctuation. The NLTK tokenizer is more robust. It tokenizes a sentence into words and punctuation.
- *Stop-words*: The NLTK stop-words from the NLTK corpus is used to remove the stop-words from the tokenized sentence.
- *Stemmer*: The Porter Stemmer is a well known stemming algorithm of NLTK. This is used to all the words including the polysemy word.

- *Database*: This includes a table that contains polysemy words along with their respective clue words.
- *Oxford Dictionary*: An API that is used to get all possible meaning of polysemy word.
- *Output module*: This module consists of meanings of polysemy word

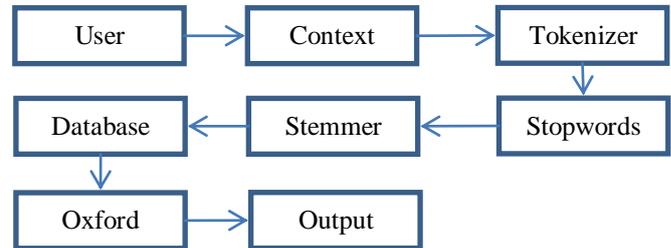


Figure 1: The Proposed Architecture

5. IMPLEMENTATION & RESULTS

The implementation details are as follows:

- First a sentence containing polysemy word is accepted from the user interface. This input text is then fed to the context module.
- The algorithm then tokenizes the sentences into words using NLTK tokenizer. The NLTK stop-words remove all the stop-words from the tokenized sentence.
- All the words including the polysemy word are stemmed using NLTK's popular stemmer known as Porter Stemmer.
- The algorithm identifies the polysemy word and consults a database containing the polysemy word along with a set of clue words for that polysemy word.
- The oxford dictionary API is used to fetch all the meanings of polysemy word irrespective of its context.
- The clue words for the polysemy word can be used as a key for context.
- Finally the oxford dictionary fetches the appropriate meaning of polysemy word based on its use in context.

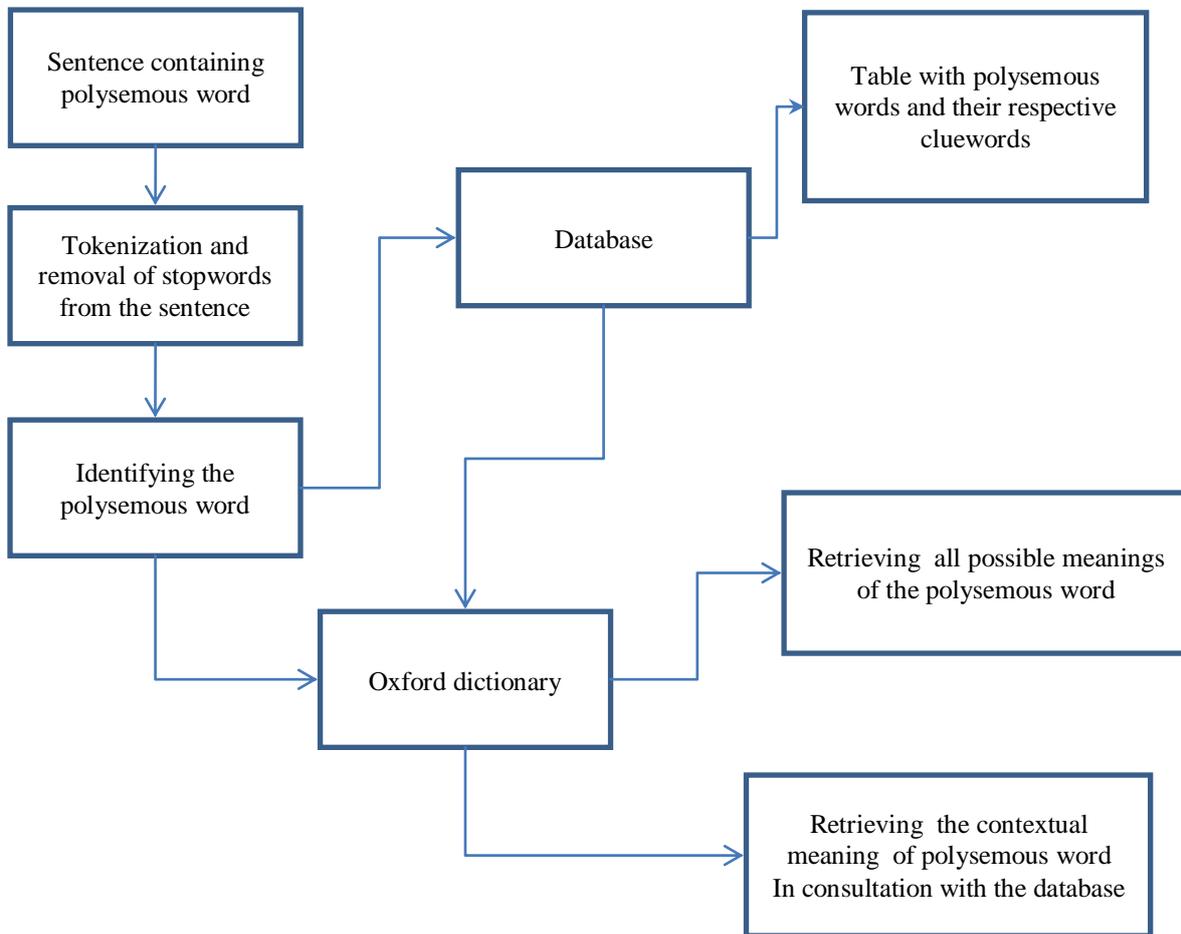


Figure 2: Flowchart Describing the Proposed Algorithm

6. CONCLUSION AND FUTURE WORKS

Identification and disambiguation of polysemy words in any text is very essential for accurate natural language processing. Semantically, a polysemy word poses challenges in language processing tasks. This paper attempts to identify and disambiguate the polysemy words in a novel way. This is possible by providing the contextual meaning of polysemy word by using the words which surround the polysemy word as clue words. In future, a comprehensive tool in identifying as well as disambiguating polysemy words can be developed. This will have great significance in many text processing tasks under machine learning and sentiment analysis.

REFERENCES

1. Hauer, B., Luan, Y., & Kondrak, G. (2019). You Shall Know the Most Frequent Sense by the Company it Keeps, 2019 IEEE 13th International Conference on Semantic Computing (ICSC). doi:10.1109/icosc.2019.8665531.
2. Lee, W. J., & Mit, E. (2011). *Word Sense Disambiguation by using domain knowledge*. 2011 International Conference on Semantic Technology and Information Retrieval. doi:10.1109/stair.2011.5995795.
3. Sachdeva, P., Verma, S., & Singh, S. K. (2014). *An improved approach to word sense disambiguation*. 2014 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT). doi:10.1109/isspit.2014.7300594
4. Lian Wang, Song, W., & Cheung, D. (n.d.). Using contextual semantics to automate the Web document search and analysis. Proceedings of the First International

- Conference on Web Information Systems Engineering. doi:10.1109/wise.2000.882845
5. Copestake, A., & Briscoe, T. (1995). Semi-productive polysemy and sense extension. *Journal of Semantics*, 12(1), 15-67.
6. Evans, V., & Green, M. (2006). *Cognitive Linguistics: An Introduction*. Edinburgh: Edinburgh University Press
7. Ingrid Lossius Falkum & Agustin Vicente. *Polysemy: Current Perspectives and Approaches*.
8. Prasad, C., & Kallimani, J. S. (2016). *A novel approach to identify polysemy words in Indian regional language*. 2016 International Conference on Electrical, Electronics, Communication, Computer and Optimization Techniques (ICEECCOT). doi:10.1109/iceeccot.2016.7955210
9. Carston, R. (2002). *Thoughts and Utterances: The Pragmatics of Explicit Communication*. Oxford: Blackwell Publishers
10. Recanati, F. (2004), *Literal Meaning*. Cambridge: Cambridge University Press.
11. Wilson, D., & Carston, R. (2007). A unitary approach to lexical pragmatics: Relevance, inference and ad hoc concepts. In N. Burton-Roberts (Ed.), *Pragmatics* (pp. 230-259). London: Palgrave
12. <https://en.wikipedia.org/wiki/Polysemy>
13. https://www.researchgate.net/publication/315458677_Polysemy
14. https://www.researchgate.net/publication/272740167_Polysemy_Current_perspectives_and_approaches
15. <http://aclweb.org/anthology/Y02-1031>