Automatic Text Standardisation by Synonym Mapping

Stella Asiimwe[1], Susan Craw[1], Nirmalie Wiratunga[1], and Bruce Taylor[2]

[1] School of Computing
[2] The Scott Sutherland School
The Robert Gordon University, Aberdeen, Scotland, UK
{sa, smc, nw}@comp.rgu.ac.uk, B.Taylor@rgu.ac.uk

Abstract. Creating a case-based reasoning system from textual sources is challenging because it requires that the text be interpreted in a meaningful way in order to create cases and interprete queries during problem-solving. However, this is only possible if a domain-specific conceptual model is available and if the different meanings that words can take, can be recognised in the text. This paper presents an unsupervised methodology for text harmonisation in order to create a standard conceptual model of the domain, and for query interpretation. The main core of this technique is the disambiguation of each polysemous word in order to represent all synonyms of a particular word, with one common word. The algorithm for Word Sense Disambiguation relies on the syntactic structure of the sentence in which the target word resides to obtain the context. It does not rely on large amounts of hand-crafted knowledge or statistical data from the corpus, but makes use of two other knowledge sources, namely WordNet and Google. Preliminary results are promising when the algorithm is tested on documents in the SmartHouse domain.

Keywords: *Word Sense Disambiguation, text harmonisation*

1 Introduction

Designing a case-based reasoning system can be challenging when the problem-solving experiences are captured as text. This is because the text has to be interpreted in a meaningful way in order to create cases that are useful for problem-solving. However, correct interpretation of text is only possible if the domain is familiar to the reader, otherwise, they will either misinterpret or, not understand, what they read. Consider a construction worker who encounters the word *mortar* whilst reading a cookery book. Unless they have knowledge of the domain, they will take *mortar* to imply a *a mixture of lime or cement with sand and water*, thus misinterpreting the word and consequently, not understanding the text. Thus, if a machine is to correctly interpret a piece of text, whether for purposes of case authoring, or interpretation of queries during problem-solving, it should have access to a conceptual model of the domain in question. The model will comprise important words and phrases in the domain and their relationships with each other. This model can be a template, or a basis for a template, on to which the texts or queries are mapped, in order to create cases or retrieve similar cases for solving new problems.

Documents describing important domain concepts are a good source of knowledge for creating conceptual models. However, textual cases by their nature are usually heterogeneous as the same words are seldom used to describe the same situations. Queries are also likely to comprise words or phrases that have the same meaning as, but are different from, those containing case knowledge. Thus it is essential to have a language model upon which the conceptual model is built, and with which texts will be interpreted during case authoring and query processing. Creation of the language model involves developing a domain-specific synonym mapper so that although a different word is used in the conceptual model, synonyms of that word will be recognised as such enabling mapping of heterogeneous texts to a standard conceptual model.

This paper focuses on creation of the language model and interpreting the texts according to that model, in order to create a harmonised knowledge source with which the conceptual model will be created. The task of modelling the language is applied to texts in the SmartHouse domain [1]. The rest of the paper is organised as follows. Related work is presented in Section 2 after which we describe the process of creating synonym mappings in Section 3. Section 4 details how the synonym mappings and a WSD module are used to harmonise the text. The WSD algorithm is evaluated in Section 5 before our concluding remarks in Section 6.

2 Related Work

In this work, each polysemous word is disambiguated in order to map synonymous words to a common token to obtain standardised text. Thus the WSD tool is the core part of this work and it is what we shall focus on to put the work in the context of wider research. Research in WSD has its roots in Weaver's work [3] where he suggested a window of words that would help put the word to be disambiguated into context and thus successfully disambiguate it. Early approaches [4, 5] to WSD were based on rule-based systems which heavily relied on large amounts of hand-crafted knowledge. The cost of manually creating the rules rendered the systems untenable for large, usable systems. Thus these approaches were only useful in restricted domains. To overcome this weakness, more recent efforts [6, 7] are based on statistical techniques which scale up more easily and can thus be applied to large corpora. One major drawback of these approaches is that the corpus may not have sufficient information to identify the appropriate word/word relationships to support disambiguation. Approaches such as that presented by Xiaobin et al. [8] overcome this shortcoming by making use of WordNet to constrain the set of possible word classes and like ours, depends on syntactic rather than statistical information. Our work is also related to that presented by Klapafttis and Manandhar [9] who use Google to find contextually related terms. The terms help in assigning the correct WordNet sense to each term under disambiguation. However, their approach depends on frequency information regarding the word to be disambiguated which makes it unattractive for sparse domains. Our method neither relies on hand-tagged data nor on statistical information of words in the document collection.

3 Creating Domain-specific Synonym Mappings

In order for a machine to make any reasonable attempt at interpreting the text it reads, it is important that it is able to recognise the different meanings of a given word or phrase. For this, it needs to have access to knowledge of the different synonyms a given word can take in the domain. A common approach to determining semantically related words is to determine their distributional similarities [10–13]. The commonality among these approaches is that they provide ranked lists of semantically related words where one would expect synonymous words to appear at the top and hyponyms and hypernyms at the lower ranks since the similarity between a word and its synonym should be higher than that between the word and say, its hypernym. Unfortunately, it is not always the case that the ranked lists reflect these different similarities. In order to ensure the reliability of synonym mappings, we use WordNet [14] to look up the synonyms of the various polysemous words in our document collection.

3.1 Creating the Clusters

We cluster words according to their synonymy relationships; words that are synonyms in a given context are clustered together. In order to determine the context, we need the word's part-of-speech and the different meanings or *senses* of each word. A CLAWS part-of-speech tagger [15] was used for this purpose.

WordNet is a lexical knowledge base where synonymous nouns, verbs, adjectives and adverbs are grouped into *synsets* where a synset represents a lexical concept. If two words with the same part-of-speech appear in the same WordNet synset, they are deemed to be synonymous in a particular context, and thus clustered together. So we build clusters of synonymous words and out of each cluster, we choose the word that appears most frequently in our texts and use it as a representative for the other words in the cluster. The representative word will substitute all the cluster words in text, during harmonisation. If more than one word in a cluster have the highest frequency, a representative is randomly chosen. The assumption here is that the most frequently used version of a word is also the most preferred in the domain. However, it is the frequency of the word in the particular part-of-speech that is considered.

Furthermore, to ensure that words from unseen text will be correctly interpreted, we include in these clusters, all other synonyms of the representative word for this sense. Thus what this process does is to select WordNet synsets that are relevant to the domain and use the most frequent words in the synsets as their representatives. These representative words will be used in place of each word in a synset, whenever the word appears in one of our documents or a query.

4 Harmonising the Text

The harmonisation task aims to remove syntactical differences between texts that are semantically equal so that the process of creating a conceptual model interprets similar texts as such. Words like "*dwelling*", "*house*", and "*home*" all mean essentially the same thing and as such, one of them could be used as a substitute for the others and still retain

the original meaning of the sentence. This results in less sparse and more cohesive data. The harmonisation task is analogous to *data cleansing* in the data mining community. The difference is that the data mining community aims to remove errors and duplication in data; we harmonise the text by using the same words to express a particular meaning, which in turn, also helps to avoid duplication.

Text harmonisation is carried out in two stages. In the first, possessive pronouns are replaced with one common article ***"the"***. For example, the sentence "*He found it difficult to open **his** door*" becomes "*He found it difficult to open **the** door*". In the second stage, each polysemous word is replaced with its representative synonym. However, the word needs to be assigned its appropriate sense in order to determine the right representative with which to replace it, if the meaning of the original sentence is to be retained. Therefore, this stage involves the resolution of syntactic and semantic ambiguity in text; problems that are common in natural language processing.

4.1 Resolving Ambiguity

Humans incrementally process the meaning of what they are reading, as they progress through a given sentence. Thus, understanding the text requires that one makes a decision as to the most likely of all the possible meanings of the text [16]. This is because the English language is highly ambiguous having words with multiple related meanings (polysemy), different words with the same spelling (homography), and words with different spellings that sound the same (homophony). Syntactic ambiguity is easy to resolve as it only requires that the word's part-of-speech be known to ensure that a word is replaced by one with the same part-of-speech. Resolving semantic ambiguity requires that the right sense of the word be obtained. The obvious difficulty with this task is that polysemous words will have more than one sense. The task therefore, is to find the meaning of the word, in the context in which it has been used, which is also the means by which humans carry out disambiguation.

A polysemous word w will appear in as many clusters as the number of the different meanings it has, as obtained from the document collection. However, our synonym mapper is based on the documents we have at hand and we can therefore not claim to have all the knowledge that is relevant to the domain, in these documents. This essentially means that there might be synsets that were not included because they were not shared by more than one word in our collection, a factor that triggers the clustering process. Therefore potentially, each word has at least one sense that was not stored. Thus our task is to determine if a given target word should be replaced by one of its possible representative synonyms and if so, to choose the correct synonym representative that will not change the meaning of the sentence. We shall refer to each of the potential representatives as a *potential representative*.

We apply the process of text harmonisation to documents in the SmartHouse domain [1] where, single words are often not meaningful, and thus concepts will comprise more short phrases than single-word terms. Thus not only do we want to retain the original meaning of the sentences, we also want the phrases that will be extracted in the process of creating the conceptual model, to be meaningful. Thus, when we replace a word with its representative, we want the short three or more word phrase in which the word appears, to still be grammatically sound. Furthermore, disambiguation must be

highly accurate if the language model is to be effective in case creation and query processing. We argue that one way to achieve high accuracy in unsupervised Word Sense Disambiguation, is to disambiguate **only** those words for which evidence to support the disambiguated version can be obtained.

4.2 Obtaining the Context

We have set out to determine the correct sense S_w of a given polysemous target word w in the context C_w, out of all the possible senses of the word as obtained from our synonym mapper. To obtain the context of the target word, we use syntactic information contained in the sentence in which it resides. We shall refer to the sentence containing the target word as the *target sentence*. The context of the target word is then obtained as follows:

1. Obtain a parse of the target sentence by using the link grammar parser [17].
2. If the target word is a noun, use the parse to extract the noun phrase in which the target word appears. If the target word is a verb, extract the verb phrase. This shall be referred to as the *target phrase*.
3. Extract the verb or noun associated with the target phrase depending on whether the target word is a noun or a verb. We shall refer to this as the *extension*.

The target phrase, the extension, and their respective parts-of-speech, all make up the context C_w that is used to derive the correct sense S_w of the target word w in the target sentence.

4.3 The Word Sense Disambiguation Algorithm

The task is to determine if a given word w should be replaced by any of its representatives, and if so, to choose the representative that will not change the meaning of the original sentence. Suppose that a polysemous word w has several potential representatives. Word Sense Disambiguation is carried out as follows:

1. Replace the target word in the target phrase with the first potential representative. Use this *potentially standardised* phrase and the extension to create a Google query.
2. Retrieve the N documents in which the query terms appear in the **same** sentences ensuring that the potential representative and the extension have been used with the same respective parts-of-speech as the target word and extension have, in the target sentence.
3. If $N \geq 1$, then the potential representative can be used to replace the target word, in the target sentence.
4. If no documents are returned, repeat the procedure with another potential representative acting as the substitute for the target word.
5. Repeat until either at least one document is retrieved or all the potential representatives have been exhausted.

The Word Sense Disambiguation process is illustrated in Figure 1. The requirement that the potential representative and the extension appear in retrieved documents with the same respective parts-of-speech as the target word and extension do in the target sentence, is a good contributing factor towards ascertaining context. The assumption is that if there is a document in which someone has used the potential representative in the same way as the target word, then the potential representative can be used in place of the target word in this sentence, without altering the meaning.

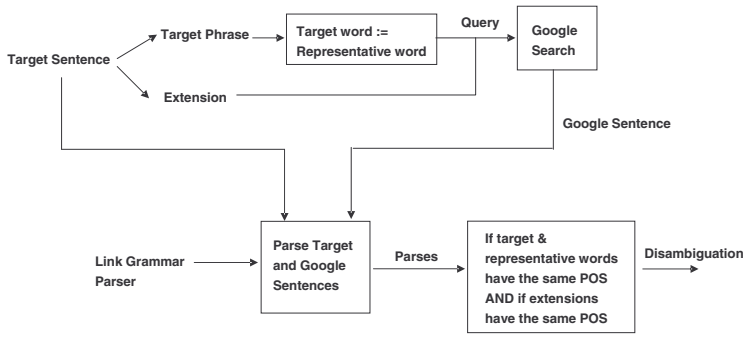

Fig. 1. The WSD Tool

The internet is used as a very large corpus where we assume that if the potentially standardised phrase is sensible, the probability of getting a document in which it appears in the same sentence as the extension and used in the same context as the target sentence, tends to 1. If no document is returned after the above operation, we have no evidence that replacing the target word with that particular potential representative will not alter the meaning of the target sentence and we thus do not replace the target word with the representative word. The level of confidence of the disambiguation is dependent on the number of returned documents in which the potentially standardised phrase is used in the same context as the target phrase. It is also dependent upon the length of the target phrase – the bigger the phrase, the more confident we are of the disambiguation. It is worth noting that although the algorithm is aimed at resolving ambiguity of single words, the use of target phrases that are not single words results in the disambiguation of some longer phrases. We shall use a simple example to illustrate the process of Word Sense Disambiguation.

Consider the sentence "*He found it difficult to open the front door*". Now, *front* has 10 senses in its noun version. In its 6th sense, *front* is synonymous with *movement*[1]. Suppose that we wish to determine if *front* should be replaced with *movement* in this sentence. Figure 2 shows a parse of the target sentence. In this case, the noun phrase *the front door* will be the target phrase and the extension will be *open*. In order to determine if *movement* is a synonym for *front* in this sentence, we replace *front* with *movement* in the target phrase changing it from *the **front** door* to *the **movement** door*. This potentially

[1] Meaning *a group of people with a common ideology.*

standardised phrase and the extension are used to create a Google query. Google returns a document on Rugby news in which the query words appear in the same sentence: "*And during the season, **the movement door** remains **open**.*" The parse of this sentence is shown in Figure 3.

```
he found.p it difficult.a to open.v the front.n door.n .

Constituent tree:

(S (NP He)
   (VP found
       (NP it)
       (ADJP difficult
             (S (VP to
                    (VP open
                        (NP the front door))))))
   .)
```

Fig. 2. A Parse of the Sentence "*He found it difficult to open the front door.*"

Examination of the parses for target and Google-returned sentences shows that, *open* in the potentially standardised phrase returned in the Google document is not used in the same context as that in our target phrase. This is evidenced by the different parts-of-speech the extension, *open*, exhibits in the different sentences. In the target phrase, the extension is used as a verb and in the Google sentence, it is used as an adjective. Since no other documents are returned where the context is the same as that of our target word, we conclude that in this instance, *front* cannot be replaced by *movement* and still retain the meaning of the original sentence.

```
and during the season.n , the movement.n door.n remains.v open.a .

Constituent tree:

(S And
   (PP during
       (NP the season))
   ,
   (S (NP the movement door)
      (VP remains
          (ADJP open)))
   .)
```

Fig. 3. A Parse of the Sentence "*And during the season, the movement door remains open.*"

5 Evaluation

We are only interested in harmonising sentences where we can obtain accurate disambiguation. Thus our focus is on the correctness of the harmonised sentences and not

necessarily on the overal effectiveness of the disambiguation tool. The Word Sense Disambiguation algorithm has been tested on documents in the SmartHouse domain [1]. Ten harmonised sentences that originally contained 14 polysemous words, were presented to a domain expert and to a native speaker of English. Since it is sometimes impossible even for humans to pick the right sense of a word from WordNet, we thought asking our two test experts to do the same would not be reasonable. Instead, we placed the representative word right next to the target word in the text and asked them whether replacing the target word with the representative word would alter the meaning of the sentence. Figure 4 shows 5 of the harmonised sentences put to our experts. The target words in the original sentences and the representative words in the harmonised sentences are shown in bold for clarity.

Original Sentence	Harmonised Sentence
She had a tendency to leave the **house**, leaving the carer behind.	She had a tendency to leave the **home**, leaving the carer behind.
She relied on **care staff** to close **her** windows.	She relied on **care worker** to close **the** windows.
He had **trouble** sleeping, constantly waking throughout the night.	He had **difficulty** sleeping, constantly waking throughout the night.
Mr. RC was at risk from fire and scalding because he was unaware of the special **dangers**.	Mr. RC was at risk from fire and scalding because he was unaware of the special **risks**.
…because of her **poor vision** relied on listening to the commentary rather than watching the action.	…because of her **poor sight** relied on listening to the commentary rather than watching the action.

Fig. 4. Sample of Original and Harmonised Sentences

All disambiguations were deemed correct by both experts i.e., replacing words with their representatives did not alter the meaning of the sentences. In 4 instances, the native English speaker thought that although the meaning of the sentence had not been altered, the representative word was as good as the word it was replacing while in the other cases, she preferred the original word. This shows that it is possible to harmonise text and not loose important information, which is what we set out to do. The sentences presented to the experts were those that our tool was able to harmonise so this was not a test of the overall performance of the Word Sense Disambiguation tool but rather, of the correctness of those words that were used to replace others during text harmonisation.

6 Conclusion

We have introduced an unsupervised approach to Word Sense Disambiguation. WordNet is used to assign each polysemous word with a sense that is applicable in the domain. The Word Sense Disambiguation algorithm makes use of syntactic information of the sentence in which a target word appears and information from Google, to assign the correct sense to the target word, in the context in which it appears. This approach has

potential applications for ontology learning and general knowledge modelling which in turn, would be useful in case authoring and query interpretation for textual CBR. The texts returned by Google provide evidence that the disambiguated phrase is plausible in the English language. Our approach is attractive because although automated, it neither relies on statistical information of the word in the document collection nor does it require the use of hand-tagged data.

References

1. S. Asiimwe, S. Craw, B. Taylor, and N. Wiratunga. Case authoring: from textual reports to knowledge-rich cases. In *Proc. of the 7th Int. Conf. on CBR*, Belfast, Northern Ireland, 2007.
2. S. Asiimwe, S. Craw, N. Wiratunga, and B. Taylor. Automatically acquiring structured case representations: The SMART way. In *Proc. of 27th SGAI Int. Conf. on AI*, 2007.
3. W. Weaver. Translation. In W. N. Locke and A. D. Boothe, editors, *Machine Translation of Languages*, pages 15–23. MIT Press, Cambridge, MA, 1949/1955. Reprinted from a memorandum written by Weaver in 1949.
4. G. Hirst. *Semantic interpretation and the resolution of ambiguity*. Cambridge University Press, New York, NY, USA, 1987.
5. S. Small and C. Rieger. Parsing and comprehending with word experts (a theory and its realization). In W. G. Lehnert and M. H. Ringle, editors, *Strategies for Natural Language Processing*, pages 89–147. Erlbaum, Hillsdale, NJ, 1982.
6. J. F. Lehman. Toward the essential nature of statistical knowledge in sense resolution. In *AAAI '94: Proc. of the twelfth national Conf. on AI (vol. 1)*, pages 734–741, Menlo Park, CA, USA, 1994. American Association for AI.
7. E. Charniak. *Statistical Language Learning*. MIT Press, Cambridge, MA, USA, 1994.
8. X. Li, S. Szpakowicz, and S. Matwin. A wordnet-based algorithm for word sense disambiguation. In *IJCAI*, pages 1368–1374, 1995.
9. I. P. Klapaftis and S. Manandhar. Google & wordnet based word sense disambiguation. *Proc. of the first workshop on learning and extending ontologies by using machine learning methods, Int. Conf. on Machine Learning, ICML-05, Bonn, Germany*, 2005.
10. L. van der Plas and G. Bouma. Syntactic contexts for finding semantically related words. In *Computational Linguistics in the Netherlands*. LOT Utrecht, 2005.
11. J. R. Curran and M. Moens. Improvements in automatic thesaurus extraction. In *Proc. of the ACL-02 workshop on Unsupervised lexical acquisition*, pages 59–66, Morristown, NJ, USA, 2002. Association for Computational Linguistics.
12. S. C. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman. Indexing by latent semantic analysis. *Journal of the American Society of Information Science*, 41(6):391–407, 1990.
13. S. Massie, N. Wiratunga, A. Donati, and Vicari. From anomaly reports to cases. In *Proc. of the 7th Int. Conf. on Case-Based Reasoning*, Belfast, Northern Ireland, 2007.
14. C. Fellbaum. *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. The MIT Press, May 1998.
15. I. Marshall. Tag selection using probabilistic methods. In R. Garside, G. Leech, and G. Sampson, editors, *The Computational Analysis of English*, pages 42–56. 1987.
16. G. B. Simpson. Lexical ambiguity and its role in models of word recognition. *Psychological Bulletin*, 96:316–340, 1984.
17. D. Sleator and D. Temperley. Parsing english with a link grammar. In *3rd Int. Workshop on Parsing Technologies.*, 1993.