A New Measure for Extracting Semantically Related Words

Yuanyong Wang and **Achim Hoffmann**
wyy@cse.unsw.edu.au
achim@cse.unsw.edu.au
School of Computer Science and Engineering
University of New South Wales
Sydney, 2052, NSW
Australia

Abstract

The identification of semantically related terms for a given word is an important problem. A number of statistical approaches have been proposed to address this problem. Most approaches draw their statistics from a large general corpus. In this paper, we propose to use specialized corpora which focus strongly on the individual words of interest. We propose to collect such corpora through targeted queries to Internet search engines. Furthermore, we introduce a new statistical measure, *Relative Frequency Ratio*,tailored specifically for such specialized corpora. We evaluated our approach by using the extracted related terms to attack the target word selection problem in machine translation. This type of indirect evaluation is conducted because a direct evaluation on the set of related terms thus extracted relies heavily on direct human involvement and is not quantitatively comparable to others' results. Our experimental results so far are very encouraging.

1 Introduction

The identification of semantically related words from texts is an important problem in natural language processing. If successfully identified, they could be used in query expansion, word sense disambiguation, as well as document classification (Tomohiko Sugimachi and Matsuo, 2003). Another application concerns the identification of new word senses in specialized languages, which are constantly evolving and, hence, no up-to-date dictionaries exist that could cover all those word senses. Many approaches, such as co-occurrence statistics based on mutual information (Church and Hanks, 1990), the Z-score (Tomohiko Sugimachi and Matsuo, 2003), have been used in the past to tackle this problem. These approaches are limited to be used only on general corpora in which large amounts of texts are collected from sources as diverse as possible. In this paper, we call this kind of corpus a General Corpus (GC). The nature of these measures (rate high co-occurrence high and rate high frequency words low) requires generality of the corpus. Generality is defined in our paper as for a corpus not being biased toward any particular domain or particular word. Mutual information is defined as follows:

$$I(w_1, w_2) = log\frac{P(w_1, w_2)}{P(w_1)P(w_2)},$$

where $P(w_i)$ is the probability of word w_i to occur in a document and $P(w_i, w_j)$ is the probability of both words w_i and w_j to occur in a document.

If a good corpus with high generality could be obtained, the co-occurrence statistics collected from the corpus could be very good and reflect accurately the tendency of semantic association among words. But there are several innate drawbacks to these GC based approaches. Firstly, generality is often very difficult to define. Secondly, acquiring a GC with good generality is a even more difficult problem. Thirdly, given that these above two problems are properly addressed the set of related terms extracted from the corpora are still limited in number. The main reason behind the third drawback is actually the generality required by these approaches. Since no words have a particularly high frequency in the corpus. It is not difficult to prove that the number of semantically related terms extracted for a given word is usually very low, especailly if we take away function words (which co-occur with any word indiscriminately). These sets of extracted related terms, if used to match the new context of the word in question, only provide very limited disambiguating power.

Based on the above analysis, another type of corpus is needed to obtain a sufficiently large set of related terms(so that they are practically useful) for any particular word. We call this type of corpus a Word Specific Corpus (WSC). It is constructed by collecting only the texts where the particular word of interest is present. We call this word the *seed word* and in formal contexts we refer to a WSC with seed word s, by $wsc(s)$. In such corpora words which occur with low frequency in a GC may

well occur with high frequency. We call this phenomenon frequency jump. Note, frequency jump of a word is a concept based on its frequency difference between a GC and a WSC where the word occurs. For example, the word "cell" is a low frequency word in GCs, but in a WSC with the seed word "tissue", it becomes a word with very high frequency. Frequency jump is very common to WSCs. In situations like this mutual information or similar measures would not properly reflect the semantic closeness between strongly related terms. Words like "cell" and the word "tissue" in the above corpus would be assigned dramatically lower mutual information value because of their high frequency. "Mutual informtion is widely known biased towards the word frequency. The tendency of mutual information does not depend on word semantics and the kinds of corpora but only on word frequency. This causes a problem in extracting the related words of a given word using an appropriate threshold value. Most of the extracted words are low frequency words and middle frequency words are rarely extracted" (Tomohiko Sugimachi and Matsuo, 2003). The Z-score measure is then proposed in the work to help extracting more middle frequency words. But from WSCs it is those high frequency words (apart from function words) that are supposed to be extracted, which makes the applicability of mutual information even Z-score worse on WSCs. Our proposed approach is designed to better measure the semantic closeness between words in WSCs and the seed word.

In summary of the above: for the set of extracted related terms to be practically useful, it has to be sufficiently large; GCs could not provide these sufficiently large sets of terms; WSCs are thus required; Mutual information and the like measures do not work well for WSCs; We propose a new approach that works well with WSCs.

The paper is organised as follows. In section 2 we theoretically present our new method for extracting related words. This is followed in section 3 by putting it into practice and showing some of the words thus extracted. In section 4, we evaluate the measure indirectly by using the extracted related words to attack the target word selection problem in machine translation. In section 5 we compare our results to those of related works. Section 6 contains the conclusions.

2 The Relative Frequency Ratio Measure

In this paper, a new measure called Relative Frequency Ratio (RFR) is proposed to extract semantically related words from WSCs. It is based on the idea that in a given context (e.g. a sentence) surrounding a seed word some words are semantically close to the seed word, others are not. For example "information" in the sentence "We have full information on all sorts of tissue paper." is not semantically as close to "tissue" as it is to "algorithm" in the sentence "The information gain is made as large as possible by this machine learning algorithm."

It is observed that in WSCs the closer a word is semantically to the seed word the larger its frequency jump (frequency increase) would be. This provides a natural measure for the semantic closeness between an arbitrary word and the seed word. The Rative Frequency Ratio, despite being based on the same spirit, is defined in a more formal way. First of all, relative frequencies for a word in both a GC and a WSC are computed. The ratio is then computed by dividing the GC relative frequency by the WSC relative frequency. In essence relative frequency ratio is a normalized version of frequency jump.

The relative frequency ratio (RFR) for a word is given by:

$$RFR_{w(wsc(s))} = \frac{f_{w(wsc(s))}/t_{wsc(s)}}{f_{w(gc)}/t_{gc}}$$

$RFR_{w(wsc(s))}$ is the ***relative frequency ratio***. $t_{wsc(s)}$ is the ***total number of word tokens*** in the corpus ***wsc(s)***. t_{gc} is the ***total number of word tokens*** in the corpus ***gc***.

For example, one of the seed words that have been tested with this approach is "tissue". A large GC in English is compiled. A WSC is also compiled with the seed word "tissue". The relative frequency for "paper" in the GC is 0.000477 and is 0.00322 in the WSC . The RFR value is 6.75. Another two words "end" and "open" that have almost same GC relative frequency as "paper" have drastically lower RFR values with respect to the WSC. The GC relative frequency and WSC relative frequency for "end" are 0.00048 and 0.000215 respectively. Its RFR value is 0.45. Similarly, the word "open" has 0.000477 and 0.000136 as its GC relative frequency and WSC relative frequency. The RFR value is 0.29. While "Paper" occurs almost 7 times more frequently in the WSC than in the GC, both "end" and "open" occur a lot less in the WSC. In this particular case, by setting a threshold of 1 for the RFR value would easily rule out the two words "open" and "end" and keep only the word "paper" as semantically close to the seed word "tissue".

	English	German
Number of Word Tokens	1,538,152	896,413
Number of Word Types	38,508	43,449

Table 1: General corpora for English and German.

3 Extraction of Semantically Related Words

3.1 Corpora

Two GCs, one English and one German general corpus, have been compiled to provide the base relative frequency statistics. GCs are compiled by issuing a set of most frequent function words as queries and extracting all the texts from the search results. This helps to avoid any possible domain bias being introduced by the compiling process because those most frequent function words are themselves not biased toward any domain. A number of WSCs in English and German have been compiled to provide the specialized relative frequency statistics for the English seed words and their German translations. Training data(WSCs) as well as testing data are collected for three English seed words "tissue", "apron" and "attack" respectively from the top n retrievals of Google and other search engines with the seed words as queries. From each document retrieved by the search engine only three sentences surrounding or containing the first occurrence of the seed word are extracted. This is mainly based on the assumption that the publisher usually tries to give as much as possible semantic information at the word's first occurrence to restrict its sense. This assumption has shown to be valid in our experiments. The data collection is summarised in Table 1 and Figure 1. The last column in Figure 1 provides some extra information about experiment results, which is better to be read together with final results summarized in Table 2 when they are discussed in Section 5.

3.2 Identifying related words

All three words have been analyzed with the RFR measure. The threshold on the measure is experimentally set to 1, which means any word with a RFR value higher than 1 would be extracted as semantically related to the seed word. In reality it might not be optimal, the threshold could vary when the sizes of the GC and the WSC are changed, or when the generality of the GC and the desired skewness of the WSC are changed. However the basic trend where semantically closer words have higher RFR values prevails.

A sample of the extracted words (word stems actually) is shown in Figure 2 for the seed word "tissue".

Word	English		German				
	Word tokens	Word types	Translations	Tokens	Word types	Sense cases (test data)	Corretly identify
Tissue	227,959	16,293	Gewebe	32,414	8,866	535	467
			Papiertaschentuch	36,842	8,634	166	132
Apron	70,968	8,102	Vorfeld	11,774	3,860	103	80
			Schutzblech	7,683	2,423	4	0
			Vorbuhne	12,696	4,122	71	0
			Schurze	7,560	2,410	137	70
Attack	103,633	9,956	Anfall	17,235	4,086	105	93
			Ubefallen	21,536	5,422	0	0
			Angriff	12,746	3,771	357	161
			attackieren	16,713	4,621	357	161
			In Angriff Nehmen	31,474	7,489	357	161

Figure 1: The table shows the statistics of our word specific corpora for the English words 'tissue', 'apron', and 'attack' and their possible German translations. Our method does not always provide a judgement for the respective word sense (target word selection). This lets the numbers of correctly identified senses appear lower than they actually are, especially when the applicability is low. A summary of the final experiment results is found in Table 2.

Word	Frequency	Word	Frequency	Word	Frequency
cell	729	study	198	cover	138
paper	442	muscle	190	bone	136
engineer	308	disease	184	develop	131
body	300	blood	181	lung	121
soft	286	structure	180	cause	117
human	278	animal	179	cancer	115
connective	259	culture	177	handkerchief	105
organ	248	toilet	169	normal	100
function	219	layer	151	bathroom	94

Figure 2: Extracted word stems of words related to the seed word 'tissue' along with their respective occurrence frequency in the word specific corpus.

An indirect evaluation approach in next section is adopted to evaluate the quality of the sets of semantically related words extracted with this measure. Semantically related words of several seed words are used to do word sense disambiguation of those seed words for machine translation between English and German.

4 Evaluation by Machine Translation

4.1 Target word selection as word sense disambiguation

In machine translation choosing the correct translation for a word is called target word selection problem. It is also a word sense disambiguation problem. In this case, word senses are defined by their distinctive translations into another language. To attack the problem some approaches have been proposed including knowledge based approach and corpus based statistical approach (Ide and Veronis, 1998; S.Sekine and J.I.Tsujii, 1995; N.Uramoto, 1995; H.A.Lee and G.C.Kim, 2002).

To evaluate the set of semantically related words extracted with RFR measure, we adopted the corpus based approach due to increasing availability of text data and the strong performance of recent corpus based statistical approaches. The experiment is conducted between English and German. We describe the experiments with respect to word sense disambiguation first rather than directly to the target word selection problem. This is because word sense disambiguation is a broader problem and our measure could be applied to it in general. So, we put this general problem before the target word selection problem.

Semantically related words (both English and German) are extracted for all three English seed words and their German translations. German words extracted for a German translation are used later as the join context of that German word. The set of semantically related English words, however, could not be used directly for word sense disambiguation. It is unknown as to which of them suggest one sense of the seed word, which suggest another. We need to convert a set of semantically related words to several sets of sense specific words. These data could then be used to match the new context of the seed word to disambiguate it.

A clustering algorithm is used to find sense specific clusters of words from the set of semantically related words. The algorithm is essentially the same as other clustering algorithms in that it attempts to find word clusters that have the strongest internal connection and to minimize the inter-cluster connections. The difference between such algorithms is often reflected by how the algorithm defines a connection. In our algorithm the connection is defined as word co-occurrence. If two words co-occur frequently enough (i.e. beyond coincidence) a connection is said to exist between them.

Surprisingly, during our experiments, we observed that the clusters found are not really sense clusters as many such clusters correspond to one sense of the seed word and many correspond another. Eventually, we come to realize that the co-occurrence statistics based clustering algorithm only goes half way to obtain sense specific clusters. Each cluster obtained should be, instead, called usage cluster. An concrete example would better explain the situation. In this example the word “tissue” is used as the seed word. A set of semantically related words are extracted with the RFR measure. From the words several clusters are obtained by the clustering algorithm. One cluster contains the words like “toilet”,“bathroom”,“roll”,“dispenser” etc, which clearly indicate the word used in the context of bathroom in the sense of toilet tissue. Another cluster contains the words like “flower”,“scissors”,“glue”,“colour”,“fold”,“cut” etc, which indicate the word being used in the context of handcrafts making in the sense of soft tissue paper as a material. These two clusters are difficult to be joined together based on co-occurrence because these two contexts rarely co-occur. In English we could say they represent different senses of the word, but in German they only have one translation. Or even in English if we take a broader view, we could say that they represent the same sense as a type of paper (in contrast to body tissues like organs). So they really correspond to word usage rather than word senses.

Unambiguously, the next step would be to join the usage clusters to form sense clusters, which is not an easy task. Different contexts (usage) rarely co-occur within a close vicinity. One fact, however, simplifies the process. Since senses of a English seed word is defined as the word’s German translations. We could bypass the English sense cluster and directly join English usage clusters under German translations of the seed word. This could be done easily with the help of a bilingual dictionary. For example, the English word “tissue” has two usage clusters aforementioned. They are used in two different contexts. In German, however, the word “Papiertaschentuch” as a translation to “tissue” is used in both contexts. If we look up words from the two English usage clusters in a bilingual dictionary, naturally many of their German translations would all occur in the German contexts of “Papiertaschentuch”. Thus we could join two English usage clusters under a German translation whenever we could match German translations of English words from both clusters to the context of that German word. The context of a German word is conveniently provided by the set of semantically related words extracted for it.

In summary, a set of semantically related words

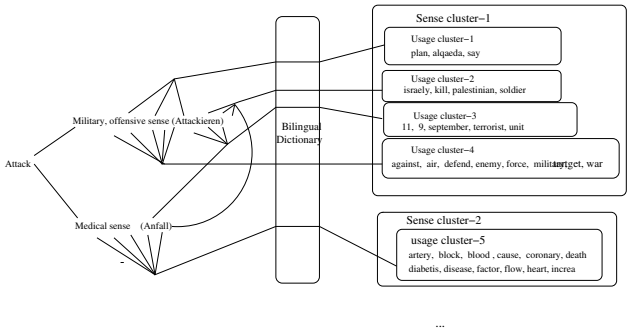

Figure 3: Usage clusters in English words are grouped around word senses based also on the German word-specific corpora.

are extracted for an English seed word X; a set of usage clusters are formed from this set; these usage clusters have to be joined under German translations of X; each usage cluster has a German representation which is the set of all possible enumerations of German translations to all the words in the cluster; this German representation overlaps with contexts of all German translations of X to different extent. The usage cluster Y would be assigned to one German translation context that has the biggest overlap with one of Y's German representations. Figure 3 shows how multiple English usage clusters could be joined under a single German translation (word). In the diagram for example, the English usage cluster2 passes through the "glass bar" of a bilingual dictionary, its German representation coming out of the dictionary has three matches with the first German translation's context, but only one match with the second German translation's context. It should be joined under the first German translation. Usage clusters thus joined under one German translation form a sense cluster that could be used to match new contexts of the seed word to disambiguate it. One sense cluster here corresponds to one German translation (word).

4.2 Testing results

Test data are collected from the Internet and are different from the training data. The sentences containing the English seed word are then labelled manually with the proper German translations of the seed word. Each test data set contains several hundred of such sentences. The sense clusters obtained with above approach are used to provide evidence for sense disambiguation alone and no other types of knowledge is used. The sense cluster that has the biggest overlap (words matched) with the new context assigns its corresponding German translation to the test sentence. This translation is compared to the correct translation manually tagged to the sentence. The results are summarized in Table 2. The 'Precision'(i.e. accuracy) column shows how often an assigned selection is correct. "Recall" indicates the percentage of cases where a judgment is made by the process.

Word	Testing data	Precision	Applicability
tissue	703 instances	97.5%	87.2%
apron	315 instances	69%	67.3%
attack	471 instances	93.7%	57.4%

Table 2: Summary of the disambiguation results. Precision is defined as the portion of correct judgements in the total number of judgments made. Applicability is defined as the portion of cases where a judgement is made in all tested cases.

5 Comparison with Related Works

There are other works that address the same problem of target word selection with different approaches. Sugimachi et al. in (Tomohiko Sugimachi and Matsuo, 2003) have used the Z-score (a refined derivative of mutual information) to extract semantically related words and form clusters from word graphs that resulted from the extraction. Their approach to the word sense disambiguation problem was evaluated qualitatively. Marquez in (Lluís Màrquez, 2000) compared five different supervised statistical approaches for WSD. They are Naive Bayes, Example Based Classifier, Winnow-based Classifier. They also investigated the effect of Boosting and Lazy Boosting. Their Lazy Boosting approach performed the best at an average of 71% accuracy on 21 selected words.

McDonald in (McDonald, 1998) used a vector distance calculation based multidimensional semantic space to calculate the closeness between alternative translations and the local context vector. Experimental results showed an accuracy around 58% at 100% recall, i.e. a judgment is made in every case. Khoen & Knight in (Koehn and Knight, 2000) used unrelated monolingual corpora in both languages together with a bilingual lexicon to build a translation model for 3830 German and 6147 English noun tokens. The probability distribution of different translations were estimated. They showed that the accuracy of their approach lies around 70% on average for a large collection of words.

Compared to these results our results are very encouraging, as our average accuracy is significantly higher. In particular, if we had used default

decisions provided in (Koehn and Knight, 2000), the recall would be much higher without substantially reducing the precision. What's important is that this machine translation application uses as its main knowledge only the set of semantically related words extracted with RFR. This (although indirectly) is sufficient as a proof to the effectiveness of the RFR measure we propose in this paper. The last thing worth mentioning is that mutual information has been used in place of RFR at early stages of the experiment but the precision rate stays at around 75 to 80% on average. Simply replacing mutual information with RFR under the exactly same framework pushes the rate up to 87% on average without compromising the applicability. One major difference made by RFR in comparison to mutual information is extraction of semantically related words with very high frequency. These high frequency words from WSCs all play vital roles in constraining the sense usage of the seed words.

6 Discussion and Conclusion

In this paper we introduced a new statistical measure RFR to extract semantically related words for a given word. The method can be applied to word sense disambiguation in general although we only showed how it could be applied to target word selection. Our experiments showed encouraging results, but because of time and resource limits it is only conducted for a small number of words so far.

The RFR measure could be used to obtain lists of domain specific words, topic specific words and basically, as long as a biased corpus could be obtained the list of words that are related to the biase could be extracted by the RFR measure. In our paper the WSCs are such corpora biased toward single words.

Some of the challenges that this measure faces are the same to those of current co-occurence based statistical approaches. One is to obtain a GC that is large enough and with good generality to provide good base statistics. But the seriousness of the problem is reduced by the fact that RFR measure is not overall sensitive to the bias unlike mutual information. If the GC is biased toward one domain, it will only affect the extraction of semantically related words for seed words in this domain. How could we better utilize the set of semantically related words is also an challenging problem. A general impression developed during the experiment of using the extracted words for WSD is that the measure often performs strongly in extracting domain or topic specific words. But word sense division does not often coincide with domain differences of the divided senses. Quite many sense divisions are based on local syntactic interaction of the word with surrounding words. This type of sense division is typical to verbs, nouns that originate from verbs and sometimes nouns with fine sense divisions. The extracted and clustered words usually do not perform well in this case. The core of the difficulty could just be the simple use of only word form co-occurrence information during the extraction and clustering. Future development of the work would be likely to focus on integrating other types of knowledge beyond word forms into the measure as well as finding of less demanding applications compared to WSD.

References

Kenneth Ward Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29.

H.A.Lee and G.C.Kim. 2002. Translation selection through source word sense disambiguation and target word selection. In *In COLONG-2002*.

Nancy Ide and Jean Veronis. 1998. Introduction to the special issue on word sense disambiguation: The state of the art. In *Computational Linguistics. Special Issue on Word Sense Disambiguation*.

Philipp Koehn and Kevin Knight. 2000. Estimating word translation probabilities from unrelated monolingual corpora using the em algorithm. In *Proceddings of the AAAI/IAAI 2000*. AAAI, Austin, Texas, USA.

Lluís Màrquez. 2000. Machine learning and natural language processing. Technical Report LSI-00-45-R, Departament de Llenguatges i Sistemes Informàtics (LSI), Universitat Politecnica de Catalunya (UPC), Barcelona, Spain.

Scott McDonald. 1998. Target word selection as proximity in semantic space. In *COLING-ACL*, pages 1496–1498.

N.Uramoto. 1995. Automatic learning of knowledge for example-based disambiguation of attachment. In *In Sixth International Conference on Theoretical and Methodological Issues in Machine Translation*, pages 295–302, KU Leuven, Belgium.

S.Sekine and J.I.Tsujii. 1995. Automatic acquisition of semantic collocation from corpora. *Machine Translation*, 10(3):218–258.

Masayuki Takeda Tomohiko Sugimachi, Akira Ishino and Fumihiro Matsuo. 2003. A method of extracting related words using standardized mutual information. In *Proceedings of 6th Discovery Science Conference*, Sapporo, Japan, October. Springer-Verlag.