Automatic Acquisition of English Topic Signatures Based on a Second Language

Xinglong Wang
Department of Informatics
University of Sussex
Brighton, BN1 9QH, UK
xw20@sussex.ac.uk

Abstract

We present a novel approach for automatically acquiring English topic signatures. Given a particular concept, or word sense, a topic signature is a set of words that tend to co-occur with it. Topic signatures can be useful in a number of Natural Language Processing (NLP) applications, such as Word Sense Disambiguation (WSD) and Text Summarisation. Our method takes advantage of the different way in which word senses are lexicalised in English and Chinese, and also exploits the large amount of Chinese text available in corpora and on the Web. We evaluated the topic signatures on a WSD task, where we trained a second-order vector co-occurrence algorithm on standard WSD datasets, with promising results.

1 Introduction

Lexical knowledge is crucial for many NLP tasks. Huge efforts and investments have been made to build repositories with different types of knowledge. Many of them have proved useful, such as WordNet (Miller et al., 1990). However, in some areas, such as WSD, manually created knowledge bases seem never to satisfy the huge requirement by supervised machine learning systems. This is the so-called knowledge acquisition bottleneck. As an alternative, automatic or semi-automatic acquisition methods have been proposed to tackle the bottleneck. For example, Agirre et al. (2001) tried to automatically extract topic signatures by querying a search engine using monosemous synonyms or other knowledge associated with a concept defined in WordNet.

The Web provides further ways of overcoming the bottleneck. Mihalcea et al. (1999) presented a method enabling automatic acquisition of sense-tagged corpora, based on WordNet and an Internet search engine. Chklovski and Mihalcea (2002) presented another interesting proposal which turns to Web users to produce sense-tagged corpora.

Another type of method, which exploits differences between languages, has shown great promise. For example, some work has been done based on the assumption that mappings of words and meanings are different in different languages. Gale et al. (1992) proposed a method which automatically produces sense-tagged data using parallel bilingual corpora. Diab and Resnik (2002) presented an unsupervised method for WSD using the same type of resource. One problem with relying on bilingual corpora for data collection is that bilingual corpora are rare, and aligned bilingual corpora are even rarer. Mining the Web for bilingual text (Resnik, 1999) is not likely to provide sufficient quantities of high quality data. Another problem is that if two languages are closely related, data for some words cannot be collected because different senses of polysemous words in one language often translate to the same word in the other.

In this paper, we present a novel approach for automatically acquiring topic signatures (see Ta-

ble 1 for an example of topic signatures), which also adopts the cross-lingual paradigm. To solve the problem of different senses not being distinguishable mentioned in the previous paragraph, we chose a language very distant to English – Chinese, since the more distant two languages are, the more likely that senses are lexicalised differently (Resnik and Yarowsky, 1999). Because our approach only uses Chinese monolingual text, we also avoid the problem of shortage of aligned bilingual corpora. We build the topic signatures by using Chinese-English and English-Chinese bilingual lexicons and a large amount of Chinese text, which can be collected either from the Web or from Chinese corpora. Since topic signatures are potentially good training data for WSD algorithms, we set up a task to disambiguate 6 words using a WSD algorithm similar to Schütze's (1998) context-group discrimination. The results show that our topic signatures are useful for WSD.

The remainder of the paper is organised as follows. Section 2 describes the process of acquisition of the topic signatures. Section 3 demonstrates the application of this resource on WSD, and presents the results of our experiments. Section 4 discusses factors that could affect the acquisition process and then we conclude in Section 5.

2 Acquisition of Topic Signatures

A topic signature is defined as: $TS = \{(t_1, w_1), ..., (t_i, w_i), ...\}$, where t_i is a term highly correlated to a target *topic* (or *concept*) with association weight w_i, which can be omitted. The steps we perform to produce the topic signatures are described below, and illustrated in Figure 1.

1. Translate an English ambiguous word w to Chinese, using an English-Chinese lexicon. Given the assumption we mentioned, each sense s_i of w maps to a distinct Chinese word[1]. At the end of this step, we have produced a set C, which consists of Chinese words $\{c_1, c_2, ..., c_n\}$, where c_i is the translation corresponding to sense s_i of w, and n is the number of senses that w has.
2. Query large Chinese corpora or/and a search engine that supports Chinese using each element in C. Then, for each c_i in C, we collect the text snippets retrieved and construct a Chinese corpus.

[1]It is also possible that the English sense maps to a set of Chinese synonyms that realise the same concept.

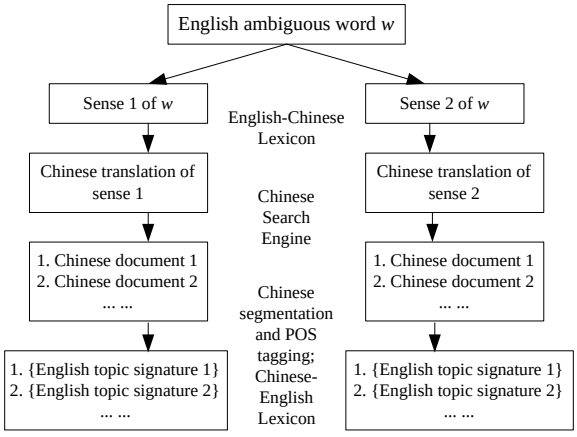

Figure 1:Process of automatic acquisition of topic signatures. For simplicity, we assume here that w has two senses.

3. Shallow process these Chinese corpora. Text segmentation and POS tagging are done in this step.
4. Either use an electronic Chinese-English lexicon to translate the Chinese corpora word by word to English, or use machine translation software to translate the whole text. In our experiments, we did the former.

The complete process is automatic, and unsupervised. At the end of this process, for each sense s_i of an ambiguous word w, we have a large set of English contexts. Each context is a topic signature, which represents topical information that tends to co-occur with sense s_i. Note that an element in our topic signatures is not necessarily a single English word. It can be a set of English words which are translations of a Chinese word c. For example, the component of a topic signature, {*vesture, clothing, clothes*}, is translated from the Chinese word 衣服. Under the assumption that the majority of c's are unambiguous, which we discuss later, we refer to elements in a topic signature as **concepts** in this paper.

Choosing an appropriate English-Chinese dictionary is the first problem we faced. The one we decided to use is the *Yahoo! Student English-Chinese On-line Dictionary*[2]. As this dictionary is designed for English learners, its sense granularity is far coarser-grained than that of WordNet. However, researchers argue that the granularity of WordNet is too fine for many applications, and some also proposed new evaluation standards. For example, Resnik and Yarowsky (1999) sug-

[2]See: http://cn.yahoo.com/dictionary/

gested that for the purpose of WSD, the different senses of a word could be determined by considering only sense distinctions that are lexicalised cross-linguistically. Our approach is in accord with their proposal, since bilingual dictionaries interpret sense distinctions crossing two languages.

For efficiency purposes, we extract our topic signatures mainly from the Mandarin portion of *the Chinese Gigaword Corpus* (CGC), produced by the LDC[3], which contains $1.3GB$ of newswire text drawn from *Xinhua* newspaper. Some Chinese translations of English word senses could be sparse, making it impossible to extract sufficient training data simply relying on CGC. In this situation, we can turn to the large amount of Chinese text on the Web. There are many good search engines and on-line databases supporting the Chinese language. After investigation, we chose *People's Daily On-line*[4], which is the website for *People's Daily*, one of the most influential newspaper in mainland China. It maintains a vast database of news stories, available to search by the public. Among other reasons, we chose this website because its articles have similar quality and coverage to those in the CGC, so that we could combine texts from these two resources to get a larger amount of topic signatures. Note that we can always turn to other sources on the Web to retrieve even more data, if needed.

For Chinese text segmentation and POS tagging[5] we adopted the freely-available software package — ICTCLAS[6]. This system includes a word segmenter, a POS tagger and an unknown-word recogniser. The claimed precision of segmentation is 97.58%, evaluated on a $1.2M$ word portion of the *People's Daily* Corpus.

To automatically translate the Chinese text back to English, we used the electronic *LDC Chinese-English Translation Lexicon Version* 3.0. An alternative was to use machine translation software, which would yield a rather different type of resource, but this is beyond the scope of this paper. Then, we filtered the topic signatures with a stop-word list, to ensure only content words are included in our final results.

One might argue that, since many Chinese words are also ambiguous, a Chinese word may have more than one English translation and thus translated concepts in topic signatures would still be ambiguous. This happens for some Chinese words, and will inevitably affect the performance of our system to some extent. A practical solution is to expand the queries with different descriptions associated with each sense of w, normally provided in a bilingual dictionary, when retrieving the Chinese text. To get an idea of the baseline performance, we did not follow this solution in our experiments.

Topic signatures for the "*financial*" sense of "*interest*"	
M	1. **rate**; 2. **bond**; 3. payment; 4. **market**; 5. **debt**; 6. dollar; 7. **bank**; 8. year; 9. loan; 10. **income**; 11. **company**; 12. inflation; 13. reserve; 14. government; 15. **economy**; 16. **stock**; 17. **fund**; 18. week; 19. security; 20. level;
A	1. {**bank**}; 2. {**loan**}; 3. {**company**, firm, corporation}; 4. {**rate**}; 5. {deposit}; 6. {**income**, revenue}; 7. {**fund**}; 8. {bonus, divident}; 9. {investment}; 10. {market}; 11. {tax, duty}; 12. {**economy**}; 13. {**debt**}; 14. {money}; 15. {saving}; 16. {profit}; 17. {**bond**}; 18. {**income**, earning}; 19. {share, **stock**}; 20. {finance, banking};

Table 1:A sample of our topic signatures. Signature M was extracted from a manually-sense-tagged corpus and A was produced by our algorithm. Words occurring in both A and M are marked in bold.

The topic signatures we acquired contain rich topical information. But they do not provide any other types of linguistic knowledge. Since they were created by word to word translation, syntactic analysis of them is not possible. Even the distances between the target ambiguous word and its context words are not reliable because of differences in word order between Chinese and English. Table 1 lists two sets of topic signatures, each containing the 20 most frequent nouns, ranked by occurrence count, that surround instances of the *financial* sense of *interest*. One set was extracted from a hand-tagged corpus (Bruce and Wiebe, 1994) and the other by our algorithm.

3 Application on WSD

To evaluate the usefulness of the topic signatures acquired, we applied them in a WSD task. We adopted an algorithm similar to Schütze's (1998)

[3]Available at: http://www.ldc.upenn.edu/Catalog/

[4]See: http://www.people.com.cn

[5]POS tagging can be omitted. We did it in our experiments purely for convenience for error analysis in the future.

[6]See: http://mtgroup.ict.ac.cn/~zhp/ICTCLAS/index.html

context-group discrimination, which determines a word sense according to the semantic similarity of contexts, computed using a second-order co-occurrence vector model. In this section, we firstly introduce our adaptation of this algorithm, and then describe the disambiguation experiments on 6 words for which a gold standard is available.

3.1 Context-Group Discrimination

We chose the so-called context-group discrimination algorithm because it disambiguates instances only relying on topical information, which happens to be what our topic signatures specialise in[7]. The original context-group discrimination is a disambiguation algorithm based on clustering. Words, contexts and senses are represented in Word Space, a high-dimensional, real-valued space in which closeness corresponds to semantic similarity. Similarity in Word Space is based on second-order co-occurrence: two tokens (or contexts) of the ambiguous word are assigned to the same sense cluster if the words they co-occur with themselves occur with similar words in a training corpus. The number of sense clusters determines sense granularity.

In our adaptation of this algorithm, we omitted the clustering step, because our data has already been sense classified according to the senses defined in the English-Chinese dictionary. In other words, our algorithm performs sense classification by using a bilingual lexicon and the level of sense granularity of the lexicon determines the sense distinctions that our system can handle: a finer-grained lexicon would enable our system to identify finer-grained senses. Also, our adaptation represents senses in Concept Space, in contrast to Word Space in the original algorithm. This is because our topic signatures are not realised in the form of words, but concepts. For example, a topic signature may consist of {*duty, tariff, customs duty*}, which represents a concept of "a government tax on imports or exports".

A vector for concept c is derived from all the close neighbours of c, where close neighbours refer to all concepts that co-occur with c in a context window. The size of the window is around 100 words. The entry for concept c' in the vector for c records the number of times that c' occurs close to c in the corpus. It is this representational vector space that we refer to as Concept Space.

In our experiments, we chose concepts that serve as dimensions of Concept Space using a frequency cut-off. We count the number of occurrences of any concepts that co-occur with the ambiguous word within a context window. The $2,500$ most frequent concepts are chosen as the dimensions of the space. Thus, the Concept Space was formed by collecting a n-by-$2,500$ matrix M, such that element m_{ij} records the number of times that concept i and j co-occur in a window, where n is the number of concept vectors that occur in the corpus. Row l of matrix M represents concept vector l.

We measure the similarity of two vectors by the cosine score:

$$corr(\vec{v}, \vec{w}) = \frac{\sum_{i=1}^{N} \vec{v_i}\vec{w_i}}{\sqrt{\sum_{i=1}^{N} \vec{v_i}^2 \sum_{i=1}^{N} \vec{w_i}^2}}$$

where $\vec{v}$ and $\vec{w}$ are vectors and N is the dimension of the vector space. The more overlap there is between the neighbours of the two words whose vectors are compared, the higher the score.

Contexts are represented as context vectors in Concept Space. A context vector is the sum of the vectors of concepts that occur in a context window. If many of the concepts in a window have a strong component for one of the topics, then the sum of the vectors, the context vector, will also have a strong component for the topic. Hence, the context vector indicates the strength of different topical or semantic components in a context.

Senses are represented as sense vectors in Concept Space. A vector of sense s_i is the sum of the vectors of contexts in which the ambiguous word realises s_i. Since our topic signatures are classified naturally according to definitions in a bilingual dictionary, calculation of the vector for sense s_i is fairly straightforward: simply sum all the vectors of the contexts associated with sense s_i.

After the training phase, we have obtained a sense vector $\vec{v_i}$ for each sense s_i of an ambiguous word w. Then, we perform the following steps to tag an occurrence t of w:

[7] Using our topic signatures as training data, other classification algorithms would also work on this WSD task.

1. Compute the context vector $\vec{c}$ for t in Concept Space by summing the vectors of the concepts in t's context. Since the basic units of the test data are words rather than concepts, we have to convert all words in the test data into concepts. A simple way to achieve this is to replace a word v with all the concepts that contain v.
2. Compute the cosine scores between all sense vectors of w and $\vec{c}$, and then assign t to the sense s_i whose sense vector $\vec{s_j}$ is closest to $\vec{c}$.

3.2 Experiments and Results

We tested our system on 6 nouns, as shown in Table 2, which also shows information on the training and test data we used in the experiments. The training sets for *motion*, *plant* and *tank* are topic signatures extracted from the CGC; whereas those for *bass*, *crane* and *palm* are obtained from both CGC and *the People's Daily On-line*. This is because the Chinese translation equivalents of senses of the latter 3 words don't occur frequently in CGC, and we had to seek more data from the Web. Where applicable, we also limited the training data of each sense to a maximum of $6,000$ instances for efficiency purposes.

Word	Sense	Training		Test		'Supervised' Baseline	Precision
bass	1. fish	418	1203	10	107	90.7%	93.5%
	2. music	825		97			
crane	1. bird	829	2301	24	95	74.7%	76.6%
	2. machine	1472		71			
motion	1. physical	6000	9265	141	201	70.1%	69.7%
	2. legal	3265		60			
palm	1. hand	852	1248	143	201	71.1%	76.1%
	2. tree	396		58			
plant	1. living	6000	12000	86	188	54.3%	70.2%
	2. factory	6000		102			
tank	1. container	6000	9346	126	201	62.7%	70.1%
	2. vehicle	3346		75			

Table 2:Sizes of the training data and the test data, baseline performance, and the results.

The test data is a binary sense-tagged corpus, the TWA Sense Tagged Data Set, manually produced by Rada Mihalcea and Li Yang (Mihalcea, 2003), from text drawn from the British National Corpus. We calculated a 'supervised' baseline from the annotated data by assigning the most frequent sense in the *test* data to all instances, although it could be argued that the baseline for unsupervised disambiguation should be computed by *randomly* assigning one of the senses to instances (e.g. it would be 50% for words with two senses).

According to our previous description, the $2,500$ most frequent concepts were selected as dimensions. The number of features in a Concept Space depends on how many unique concepts actually occur in the training sets. Larger amounts of training data tend to yield a larger set of features. At the end of the training stage, for each sense, a sense vector was produced. Then we lemmatised the test data and extracted a set of context vectors for all instances in the same way. For each instance in the test data, the cosine scores between its context vector and all possible sense vectors acquired through training were calculated and compared, and then the sense scoring the highest was allocated to the instance.

The results of the experiments are also given in Table 2 (last column). Using our topic signatures, we obtained good results: the accuracy for all words exceeds the supervised baseline, except for *motion* which approaches it. The Chinese translations for *motion* are also ambiguous, which might be the reason that our WSD system performed less well on this word. However, as we mentioned, to avoid this problem, we could have expanded *motion*'s Chinese translations, using their Chinese monosemous synonyms, when we query the Chinese corpus or the Web. Considering our system is unsupervised, the results are very promising. An indicative comparison might be with the work of Mihalcea (2003), who with a very different approach achieved similar performance on the same test data.

4 Discussion

Although these results are promising, higher quality topic signatures would probably yield better results in our WSD experiments. There are a number of factors that could affect the acquisition process, which determines the quality of this resource. Firstly, since the translation was achieved by looking up in a bilingual dictionary, the deficiencies of the dictionary could cause problems. For example, the *LDC Chinese-English Lexicon* we used is not up to date, for example, lacking entries for words such as 手机 (mobile phone), 互联网 (the Internet), etc. This defect makes our WSD algorithm unable to use the possibly strong topical information contained in those words. Secondly, errors generated during Chinese segmentation could affect the distributions of words. For example, a

Chinese string ABC may be segmented as either $A + BC$ or $AB + C$; assuming the former is correct whereas $AB + C$ was produced by the segmenter, distributions of words A, AB, BC, and C are all affected accordingly. Other factors such as cultural differences reflected in the different languages could also affect the results of this knowledge acquisition process.

In our experiments, we adopted Chinese as a source language to retrieve English topic signatures. Nevertheless, our technique should also work on other distant language pairs, as long as there are existing bilingual lexicons and large monolingual corpora for the languages used. For example, one should be able to build French topic signatures using Chinese text, or Spanish topic signatures from Japanese text. In particular cases, where one only cares about translation ambiguity, this technique can work on any language pair.

5 Conclusion and Future Work

We presented a novel method for acquiring English topic signatures from large quantities of Chinese text and English-Chinese and Chinese-English bilingual dictionaries. The topic signatures we acquired are a new type of resource, which can be useful in a number of NLP applications. Experimental results have shown its application to WSD is promising and the performance is competitive with other unsupervised algorithms. We intend to carry out more extensive evaluation to further explore this new resource's properties and potential.

Acknowledgements

This research is funded by EU IST-2001-34460 project MEANING: Developing Multilingual Web-Scale Language Technologies, and by the Department of Informatics at Sussex University. I am very grateful to Dr John Carroll, my supervisor, for his continual help and encouragement.

References

Eneko Agirre, Olatz Ansa, David Martinez, and Eduard Hovy. 2001. Enriching WordNet concepts with topic signatures. In *Proceedings of the NAACL workshop on WordNet and Other Lexical Resources: Applications, Extensions and Customizations*. Pittsburgh, USA.

Rebecca Bruce and Janyce Wiebe. 1994. Word-sense disambiguation using decomposable models. In *Proceedings of the* 32^{nd} *Annual Meeting of the Association for Computational Linguistics*, pages 139–146.

Timothy Chklovski and Rada Mihalcea. 2002. Building a sense tagged corpus with open mind word expert. In *Proceedings of the ACL 2002 Workshop on "Word Sense Disambiguation Recent Successes and Future Directions"*. Philadelphia, USA.

Mona Diab and Philip Resnik. 2002. An unsupervised method for word sense tagging using parallel corpora. In *Proceedings of the* 40^{th} *Anniversary Meeting of the Association for Computational Linguistics (ACL-02)*. Philadelphia, USA.

William A. Gale, Kenneth W. Church, and David Yarowsky. 1992. Using bilingual materials to develop word sense disambiguation methods. In *Proceedings of the International Conference on Theoretical and Methodological Issues in Machine Translation*, pages 101–112.

Rada Mihalcea and Dan I. Moldovan. 1999. An automatic method for generating sense tagged corpora. In *Proceedings of the* 16^{th} *Conference of the American Association of Artificial Intelligence*.

Rada Mihalcea. 2003. The role of non-ambiguous words in natural language disambiguation. In *Proceedings of the Conference on Recent Advances in Natural Language Processing, RANLP 2003*. Borovetz, Bulgaria.

George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J. Miller. 1990. Introduction to WordNet: An on-line lexical database. *Journal of Lexicography*, 3(4):235–244.

Philip Resnik and David Yarowsky. 1999. Distinguishing systems and distinguishing senses: New evaluation methods for word sense disambiguation. *Natural Language Engineering*, 5(2):113–133.

Philip Resnik. 1999. Mining the Web for bilingual text. In *Proceedings of the* 37^{th} *Annual Meeting of the Association for Computational Linguistics*.

Hinrich Schütze. 1998. Automatic word sense discrimination. *Computational Linguistics*, 24(1):97–123.