

Creating a Bilingual Ontology:

A Corpus-Based Approach for Aligning WordNet and HowNet

‡ Marine Carpuat
eemarine@ust.hk

† Grace Ngai
grace@weniwen.com

Pascale Fung † ‡
pascale@ee.ust.hk

Kenneth W. Church*
kwc@research.att.com

† Weniwen Technologies
Clear Water Bay
Hong Kong

‡ Human Language Technology Center
HKUST
Clear Water Bay, Hong Kong

* AT&T Shannon Laboratory
Florham Park, NJ 07932
USA

Abstract

The growing importance of multilingual information retrieval and machine translation has made multilingual ontologies an extremely valuable resource. Since the construction of an ontology from scratch is a very expensive and time consuming undertaking, it is attractive to explore ways of automatically aligning monolingual ontologies which already exist.

This paper presents a language-independent, corpus-based method that borrows from techniques used in information retrieval and machine translation, for creating a bilingual ontology by aligning WordNet with an existing Chinese ontology called HowNet. We will present results to show that our method is capable of efficiently aligning ontologies with very different structures, as well as ontologies from languages that are very different from each other.

1 Introduction

The increasing popularity of the web and the growing inter-connectivity of the world have added importance to the fields of machine translation and multilingual information retrieval. This has increased the need for multilingual ontologies. The Euro WordNet (Vossen et al., 1997), which links together languages such as Dutch, Italian, Spanish, German, French, Czech and Estonian with the American English WordNet (Miller et al., 1990), is an example of such a multilingual ontology.

Since ontology building is a very time-consuming and costly process, the ideal method of building a multilingual ontology would be to align, or link, existing monolingual ontologies.

Although such a task, at a first blush, may seem far from challenging, there are many reasons why it is less trivial than it seems. A manual alignment would take too long and be extremely expensive; while automatic alignment would run into many of the same problems faced by machine translation, among the more serious of them being the disambiguation problem.

2 Motivation and Challenge

In this paper, we propose a language independent, corpus-based method for automatically creating a bilingual ontology from two existing ontologies. Such bilingual ontologies are extremely useful for applications such as cross-lingual information retrieval and machine translation; however, current methods for construction of these ontologies rely on large amounts of human labour. There has been some work in ontology alignment (Palmer & Wu, 1999; Dorr et al., 2000); however, these methods usually mainly utilize the structural information within the ontologies and therefore may not be applicable to ontologies that have vastly differing structures.

Our proposal is to create a bilingual Chinese-English ontology by linking the American English WordNet¹ and Simplified Chinese HowNet together by their most basic concepts: the WordNet synset and the HowNet definition. The languages involved, Chinese and English, are very different from each other; and these two ontologies are very different in their structures and design philosophies, making this an excellent test for the portability and robustness of our algorithm.

¹ WordNet 1.6 was used in all our experiments.

The ready availability of bilingual dictionaries may make this seem like an easy task, but the truth of the matter is that the task is made far from trivial by the fact that:

1. The structure of HowNet and WordNet are vastly different, as will be described later, making a surface structural alignment impossible; and
2. Many polysemous words (words with more than one sense description) have multiple translations depending on the sense, making it necessary to tackle the problem of word sense disambiguation, a complicated problem.

Our method borrows from techniques found in Information Retrieval and Statistical Machine Translation. The rest of this section will first describe the WordNet and HowNet structures, and go on to give a thorough description of the algorithm.

2.1 Ontologies: WordNet and HowNet

WordNet (Miller et al., 1990) is an electronic lexical database in which English nouns, verbs, adjectives and adverbs are arranged in synonym sets, with different relationships linking the different sets.

Like WordNet, HowNet (Dong, 1988) is also an electronic lexical database for words, which are mostly in Chinese (some English technical terms such as “ASCII” are included). HowNet and WordNet, however, differ greatly in their structure, coverage and sense granularity. WordNet aims to differentiate word senses from each other through the use of synonym sets, or synsets. These synsets are constructed by gathering synonymous word senses into nodes, which are constructed such as to allow a user to easily distinguish between the different senses of a word. For example, given the noun “address”, the construction of the synsets {address, computer address}, and {address, speech} should be enough to assist a user in distinguishing between the different senses of the word.² The synsets are then in turn linked to other synsets through hierarchical relations such as hypernyms, hyponyms, holonyms and meronyms. A total of 109,377 synsets are defined.

The meaning representation in WordNet thus follows what is commonly referred to as a differential theory of lexical semantics (Miller et al., 1990); and in contrast, HowNet takes a constructive approach to building a lexical hierarchy. At the most atomic level is a set of almost 1500 basic definitions, or sememes, such as “human”, or “aValue” (attribute-value). A total of 16,788 word concepts, or definitions, are composed of subsets of these sememes, sometimes accompanied with “pointers” that express certain kinds of relations.

For example, in HowNet, the word “疤” (scar) is associated with the definition (the English glosses are part of the original HowNet definitions):

{ trace|痕, #disease|疾病, #wounded|受伤 }

where “#” is a pointer that denotes a “relevance” relation.

Unlike WordNet, synsets are not explicitly defined in HowNet. However, word synonyms can still be found by looking for words with identical definitions. For example, the words “发烧友” (fanatic) and “爱好者” (enthusiast/hobbyist) are both associated with the definition:

{ human|人, *FondOf|喜欢, #WhileAway|消闲 }

where “*” denotes a relation of “agent or instrument of an event”.

Relationships other than synonyms can also be extracted from the definitions. For example, the word “爱好” (hobby) has the definition:

{ fact|事情, \$FondOf|喜欢, #WhileAway|消闲 }

where “\$” denotes a relation of “patient, target, possession or content of an event”.

A detailed WordNet-HowNet structural comparison can be found in Wong & Fung (2002).

2.2 Word Sense Disambiguation

The biggest initial problem encountered was that of polysemous words and multiple sense translations. For example, the word “crane”, which has two sense definitions in WordNet, has at least two Chinese translations. The first of the two WordNet senses, “a piece of heavy machinery” is

² Example taken from (Miller et al., 1990)

usually translated as “起重机”, while the common translation for the second sense, “a large wading bird”, is “鹤”.

To get a sense of the severity of this problem, we conducted a simple baseline experiment:

1. Pick 2000 HowNet definitions at random and extract all the corresponding word entries.
2. Translate each of these words to English.
3. Associate each of these English words with its corresponding entry in WordNet.

Average Number of HowNet Entries per Definition	5.4
Average Number of WordNet Synsets per Definition	8.1

Table 1: Results of Baseline Experiment

Table 1 shows the results of the baseline experiment. The 2000 randomly chosen HowNet definitions each contain, on average, over 5 words, or entries. Furthermore, our simple alignment algorithm aligns each definition to an average of 8 WordNet senses.

To create a finer-grained mapping, we took a similar approach to the Definition Match Algorithm (Knight & Luk, 1994), which compares words according to their contexts from example sentences and definitions found in a dictionary. Our approach also uses word contexts, but instead of using dictionary definitions, we extract word contexts from a large bilingual corpus.

2.3 An Information Retrieval Approach to Bilingual Dictionary Extraction

In order to be able to compare words using their contexts, we need to use a method that will allow cross-lingual comparison of words and their contexts.

Fung & Lo (1998) developed an information retrieval-like method designed to allow comparison of word contexts across languages, and across corpora that need not be parallel. Given two languages, say English and Chinese, the algorithm consists of the following steps:

1. Define a list of t seed words in both languages. The set of seed words from one language is a direct translation from those in the other language.
2. For each of the other words w in the corpus, construct a context vector v relative to the seed words from the corresponding language (e.g. for an English word, consider only English seed words, etc.), such that v_i , the i^{th} element of v , is defined as:

$$v_i = TF_{iw} \times IDF_i$$

where TF_{iw} = term frequency (number of occurrences) of seedword i in the context of w

$$IDF_i = -\log \frac{\max(n)}{n_i}$$

where n_i = total number of occurrences of seedword i in the corpus

$\max(n)$ = maximum frequency of any seedword in the corpus

3. Given a pair of words from the two languages — for example, an English word w_e and a Chinese word w_c — the similarity between them is the distance between their context vectors. We use the S3 measure from Fung & Lo (1998), which is a combination of the cosine similarity and the dice coefficient:

$$similarity(w_e, w_c) = \frac{\sum_{i=1}^t (w_{ic} \times w_{ie})}{\sqrt{\sum_{i=1}^t w_{ic}^2 \times \sum_{i=1}^t w_{ie}^2}} \times \frac{2 \sum_{i=1}^t (w_{ic} \times w_{ie})}{\sum_{i=1}^t w_{ic}^2 + \sum_{i=1}^t w_{ie}^2}$$

where

$$w_{ic} = TF_{ic} \times IDF_i$$

$$w_{ie} = TF_{ie} \times IDF_i$$

2.4 Using Synsets for Word Sense Disambiguation

The context vector similarity has been shown to be effective at extracting bilingual word translation pairs. What we are interested in, however, is, assuming that we already have a list of candidate

translations, the alignment of the proper translation pair to the correct word sense. Our method uses the similarity score and the synset information to aid in word sense disambiguation.

1. Given a HowNet definition d , first extract its associated set of Chinese words and English translations.
2. For each word from the English translations, find the WordNet synsets that it belongs to.
3. For each of these candidate WordNet synsets s ,
 - 3.1. If s contains only a single word ($|s| = 1$), expand it by adding words from its direct hyperset.
 - 3.2. Define:

$$similarity(d, s) = \frac{\sum_{w_e \in s} \sum_{w_c \in d} similarity(w_e, w_c)}{\sum_{w \in s} appears(w)}$$

$$\text{where } appears(w) = \begin{cases} 1 & \text{if } n_w > 0 \\ 0 & \text{otherwise} \end{cases}$$

The candidate WordNet synsets are then ranked according to their similarity with the Chinese HowNet definition. The alignment “winner” is defined as the highest-ranking WordNet synset.

Step 3.1 deserves some more explanation. In the course of our investigation, we noticed that there were many WordNet synsets that contained only one word, which correspond to word senses for which a reasonable synonym does not exist. (WordNet uses a short gloss to help the user to distinguish between such senses.) For example, the noun “support” has 11 senses defined in WordNet, of which 5 are single-word synsets. The semantic meaning of these synsets range from “aid” to “supporting structure”. If our method relied solely on words from the synset, we would be unable to distinguish between these senses, which would be a major problem indeed. Therefore, words from the hyperset – the set of hypernyms of the current word – are included to aid in defining the meaning.

3 Experiments

The bilingual data in our experiments was taken from the English-Chinese Hong Kong News corpus, which comprises of almost 18,500 aligned article pairs (totalling over 6 million words on the English side) from news documents released between 1997 and 2000. On the Chinese side, the corpus was word segmented with a simple greedy maximum-forward-match algorithm, using the entire HowNet vocabulary as a lexicon. The seed words list for the context vector construction was extracted by taking the monosemous words from WordNet and throwing out all those that had more than one translation in Chinese.

It must be noted that even though the Hong Kong News Corpus is roughly parallel, our method does not restrict us to parallel corpora. Indeed, any large bilingual corpora will work for our method.

3.1 Overall Results

Table 2 shows the highest scoring alignments. For each HowNet definition, the highest scoring WordNet synset that was aligned to it, and the corresponding alignment score are shown. With a few exceptions, it can be seen that most of the time, our method is successful at doing a reasonable alignment of WordNet synsets to align to HowNet definitions.

The “BeNot|非” and “BeGood|良态” definitions deserve some further explanation. It is difficult to understand how “BeNot” aligns to a WordNet synset of {name, identify}. The HowNet entries that carry this definition, however, mostly have a rough meaning of “misrecognize”, or “mistaken identity”, hence the bizarre-seeming WordNet alignment result. Likewise, the entries for “BeGood” generally carry the meaning of being “in a good or desirable state”, which leads to “BeGood” aligning to the synset {state}.

HowNet Definition	Top Aligned WordNet Synset(s)	Score
human 人, #occupation 职位, employee 员	{employee, worker}	0.002456
BeNot 非	{name, identify}	0.002311
human 人, unable 庸, undesired 莠	{master, original}	0.0007193
BeRecovered 复原, StateIni=alive 活着	{revive}	0.0004365
image 图像, \$carve 雕刻	{sculpture}	0.0003106
AlterForm 变形状	{top, pinch}	0.0001777
aValue 属性值, rank 等级, elementary 初	{elementary, primary}	0.0001083
AimAt 定向	{calculate, aim, direct}	8.958×10^{-5}
attribute 属性, pattern 样式, physical 物质	{form, word form}	4.859×10^{-5}
break 折断	{break}	4.624×10^{-5}
pay 付, possession=money 货币	{pay}	3.769×10^{-5}
BeGood 良态	{state}	3.739×10^{-5}
BeOpposite 对立	{confront}	1.460×10^{-5}
donate 捐, possession=money 货币	{subscription}	1.094×10^{-5}
HoldWithHand 搀扶	{pass, hand, reach, pass on, turn over, give}	4.9565×10^{-6}
AmountTo 总计, means=CauseToBe 使之是	{convert, change over}	2.557×10^{-6}
time 时间, @rest 休息, education 教育	{break, pause, interruption}	2.173×10^{-6}
Avalue 属性值, form 形状, even 匀	{even}	1.549×10^{-6}
BeBad 衰变	{die, decease, perish, go, exit, pass away, expire}	1.792×10^{-7}
AlterLocation 变空间位置	{exchange, change, interchange}	1.4333×10^{-7}

Table 2: Top Ranking Alignments of HowNet definitions to WordNet Synsets. (Words enclosed in curly braces belong to the same synset)

3.2 Individual Examples

The previous subsection gave an overview of the results. In order to have a better idea of the method, we will consider individual examples in this subsection.

3.2.1 Example 1: Sculpture, or School Principal?

We first consider the HowNet definition

{ image|图像, \$carve|雕刻 }

which includes the words “版刻” (carving), “半身像”(bust), “雕刻”(carving), “雕塑” (sculpture), “雕像” (statue), “头像” (head), “浮雕” (relief sculpture) and “泥塑”(clay sculpture), among others. Using the words from the English translations directly, we have a mapping to 77 WordNet synsets. However, our algorithm imposes the following partial ordering on the synsets:

WordNet Synset	Similarity with HowNet definition
{sculpture} (sense 1 of sculpture)	0.0003106
{principal, school principal, head teacher, head} (sense 13 of head)	0.0003091
{mud, clay} (sense 2 of clay)	0.0003072

The WordNet synset that best fits the HowNet definition is “sculpture”, which is appropriate given that all the various objects defined by the HowNet definition are various types of sculptures.

3.2.2 Example 2: A Shape, or the Flow of Traffic?

As another example of the difficulty of our task, and to demonstrate the success of our method in distinguishing between different word senses, we consider another HowNet definition:

{ 属性 (attribute), 样式 (pattern), &物质 (physical) }

which includes the following words: “式” (pattern, form) and “式样” (model, style). The English words map to a total of 38 senses in WordNet.

Our method imposes the following partial ordering on the senses:

WordNet Synset	Similarity
{form, shape, pattern} (sense 3 of form)	3.739×10^{-5}
{form} (sense 8 of form) hypernyms: document, written document, papers	3.725×10^{-5}
{traffic pattern, approach pattern} (sense 7 of pattern)	3.680×10^{-5}

As before, the most appropriate sense is assigned the highest weight and ranking.

3.2.3 Example 3: An Interruption, or a Fracture?

Another good example of the strength of our algorithm can be demonstrated by the HowNet definition {time|时间, @rest|休息, education|教育}

This definition has only one entry: “课间” (break).

WordNet Synset	Similarity
Pause, intermission, break, interruption, suspension (sense 7 of break)	1.43×10^{-7}
fault, geological fault, fault line, fracture, break (sense 2 of break)	1.03×10^{-7}

Since “break”, in this sense usage, carries the meaning similar to that of “recess” or “term break”, our method correctly ranks the WordNet sense corresponding to {pause, intermission, break, etc} ahead of the synset that corresponds to geological faults.

3.2.4 A Problematic Example: AlterForm|变形状 and “break”

The HowNet definition {AlterForm|变形状} and the WordNet synset {break} provide challenging cases for our method. The set of words in the HowNet definition are very diverse and include entries such as 结 (congeal), 结块 (curdle/agglomerate) and 碎 (break into pieces). The WordNet synset {break}, on the other hand, is very finely divided with 63 total senses, with 34 of them being single-word synsets. Several senses of “break”, with the following partial ranking, were among the possible alignment candidates picked for this HowNet sense:

WordNet Synset	Similarity
{break} (sense 25 of break) No hypernyms	4.6234×10^{-5}
{break, break off, snap off} (sense 45 of break)	4.522×10^{-5}
{break} (sense 31 of break) Hypernyms: cancel, call off	4.468×10^{-5}
{break} (sense 60 of break) Hypernyms: decrease, diminish, lessen, fall	4.439×10^{-5}
Bankrupt, ruin, break (sense 35 of break)	4.435×10^{-5}

Sense 25 of “break” carries the meaning of “cause to give up” (as in: “to break oneself of a bad habit”). A single-word synset without any hypernyms, it creates a difficult case to handle due to the lack of other semantic clues, and indeed, our method wrongly gives it the top rank. It is encouraging, however, that sense 45 of break, which is one of the most semantically similar senses, is given the second-highest relative ranking; and senses 31 and 60, which are also single-word synsets, are given reasonable relative rankings, which would not have been possible without hyperset generalization.

3.2.5 WordNet to HowNet mappings

In addition to the above examples, the reverse mapping of WordNet synsets to HowNet definitions can also demonstrate the capabilities of our method. As an example, we consider the word “board”, which is divided into 9 WordNet senses, ranging from “plank” to “dining table” to “committee”.

When translated to Chinese, the synonyms for “board” in its various senses produce a total of 537 HowNet entries, distributed among 39 HowNet definitions. Table 3 shows, for each WordNet synset, the highest-ranking HowNet definition that was aligned to it. (The WordNet gloss is included for synsets that contain only one word for the reader’s convenience.)

WordNet Senses for "board"		HowNet Definition
1	board (a committee having supervisory powers)	display 展示
2	board (a flat piece of material designed for a special purpose)	control 控制
3	board, plank	shape 物形,flat 扁,surfacial 面
4	display panel, display board, board	display 展示 show 表演物
5	board, gameboard	display 展示 show 表演物
6	board, table	shape 物形,flat 扁,surfacial 面 wood 木
7	control panel, instrument panel, control board	control 控制
8	circuit board, circuit card	part 部件,%implement 器具
9	dining table, board	shape 物形,flat 扁,surfacial 面

Table 3: HowNet Definitions for various senses of the word "board"

In general our method does find reasonable HowNet definitions for each synset. For example, it correctly identifies "circuit board" as an "implemented part". Where it does not find the best alignment, for example, for the two senses of "table" and "dining table", it does give reasonable descriptions: a "table" is reasonably described as having a shape that is "flat" and "surficial", and furthermore is commonly made of "wood".

An important thing to point out about this example is the problem caused by data scarcity. For the first sense of the word "board", the correct HowNet definition is {institution|机构, *manage|管理, commercial|商}, which contains only one entry: the word "董事会" (board of directors, or simply, "board"). This word did occur in the Hong Kong Daily News corpus used in our experiments, but due to the sparseness of the seedwords in our corpus, its resulting similarity scores with the WordNet synsets were negligible.

4 Analysis

The examples presented in the previous section clearly show that in many cases, our method succeeds in finding a correct ordering of the WordNet senses that are candidates for alignment to a HowNet definition. However, there are some problematic situations that need to be addressed before a full HowNet-WordNet mapping can be achieved:

- One significant problem is the difference in sense granularity between WordNet and HowNet. For example, the HowNet definition "牲畜" (livestock) includes entries as diverse as "狗" (dog), "牛" (cattle), "兔" (rabbit), while WordNet allocates these three words to different synsets. Therefore, a 1-to-1 mapping from all HowNet definitions to WordNet synsets does not exist.
- The seed word coverage that we are achieving with the HK Daily News corpus is also a matter of concern. As our seed words, we picked monosemous words in WordNet which had only one Chinese translation in our dictionary. This resulted in a list of very precise translations, but unfortunately, the words thus extracted also tend to be rare words, resulting in context vectors that had a lot of blank fields. To address this concern, and also to test the robustness of our approach, we plan to further experiment on comparable, rather than parallel, corpora. Since comparable corpora tend to be more plentiful than parallel corpora, we hope that the introduction of extra noise will be more than compensated for by the larger amount of available data.
- Another major problem encountered was that of non-compositional compounds, or NCCs. NCCs are word phrases whose meanings are "a matter of convention and cannot be synthesized from the meanings of their space-delimited components" (Melamed, 1997). Examples include phrases such as "floppy disk" and "hot dog". At the present, our method considers only singleton words and does not take such compounds into account, though it is intuitive that these compounds should not be broken up. We are considering using techniques borrowed from Chinese word segmentation, together with a large dictionary, to join words in these compounds before running our method on it.

- Other concerns include the IR-like technique used to tackle this problem, which also leads to some problems. Measures such as TF and IDF do not take into account the syntactic function of a word, causing nouns and verb forms of the same word to have the same weight, which is likely suboptimal. Furthermore, techniques such as stemming should also be included into the method, which would likely be able to capture the way a word is used more accurately.

Even though these are real concerns that should be addressed, none of them appear to be due to a fundamental flaw in our algorithm, and some of them (for example, the corpus problem and the stemming issue) are easily surmounted. Our method still succeeds in the partial aligning of two ontologies, which were constructed using very different approaches and philosophies and therefore differ vastly in their structure.

5 Related Work

The previous work that directly led to our research has been described in the earlier sections. This section will describe related work that targets the same problem.

There has been some interest in aligning ontologies. Dorr et al. (2000) and Palmer & Wu (1995) took a structural approach to this problem. They focused on HowNet verbs and used thematic-role information, which denotes the contexts in which a particular verb may occur. The HowNet thematic-role specifications are mapped to word classes in an existing classification of English verbs called EVCA (Levin, 1993), whose structure is similar to that of the verb classes in HowNet. These mappings are then used to align English EVCA verbs to Chinese HowNet verbs. In Japanese, Asanoma (2001) used structural link information to align nouns from WordNet to a pre-existing Japanese ontology called Goi-Taikei via the Japanese WordNet (Hayashi 1999), which was constructed by manually translating a subset of WordNet nouns.

There has also been a lot of interest in automatic bilingual word alignment and dictionary induction. The IBM Candide project (Brown et al., 1990) used statistical data to align words in sentence pairs from parallel corpora in an unsupervised fashion through the EM algorithm. Church (1993) used character frequencies to align words in a parallel corpus, and Fung & Church (1994) used seed words to align parallel texts and extract bilingual word translation pairs.

Word sense disambiguation is also a difficult and much-studied subject, and many approaches have been tried on the problem. Among the numerous efforts are those of Gale, Church & Yarowsky (1992) and Schuetze (1992), who applied vector space models and similarity measures to the problem; Yarowsky (1995) used decision lists, and Kikui (1999), who used word in non-parallel bilingual corpora to resolve translation ambiguity in an unsupervised fashion.

6 Conclusions and Future Work

In this paper, we present a language-independent, corpus-based method to align definitions from the Chinese HowNet with synsets from the English WordNet. Unlike many other ontology alignment techniques, our method does not take into account the structure of the ontology. Our method uses techniques borrowed from machine translation and information retrieval to calculate similarity scores between HowNet definitions and WordNet synsets; the alignment then relies on these scores to find the most optimal alignment between the ontologies. Since our method does not make any assumptions about the structure of the ontology, or use any but the most basic structural information, it makes it possible to align ontologies with vastly different structures.

We show that our method is promising in its ability to produce a reasonably good mapping from HowNet definitions to WordNet synsets. There are some problematic situations that prevent us from achieving a full mapping at this point, but they do not appear to be serious or insurmountable.

We plan to expand on this work in the future by addressing the concerns raised in the analysis section and producing a full alignment from HowNet to WordNet. We also intend to expand our algorithm to possibly integrate more structural information. Finally, our goal is to examine the use of the aligned ontology in applications such as cross-lingual information retrieval and machine translation.

Acknowledgements

The authors would like to thank researchers at Weniwen Technologies — Ping-Wai Wong for his help and explanations on HowNet construction and structure; and Chi-Shun Cheung and Chi-Yuen Ma for their assistance in corpora and dictionary preparation.

References

- N. Asanoma. Alignment of Ontologies: WordNet and Goi-Taikai. In *Workshop on WordNet and Other Lexical Resources: Applications, Extensions and Customizations*. Pittsburgh, Pennsylvania, 2001.
- P.F. Brown, J. Cocke, S.A. Della Pietra, V.J. Della Pietra, F. Jelinek, J.D. Lafferty, R.L. Mercer and P. Roosin. A Statistical Approach to Machine Translation. *Computational Linguistics*, 16:79—85, 1990.
- K. Church. Char_align: A Program for Aligning Parallel Texts at the Character Level. In *Proceedings of the 31st Annual Conference of the Association for Computational Linguistics*, pp 1—8, Columbus, Ohio, 1993.
- Z. Dong. Knowledge Description: What, How and Who? In *Proceedings of International Symposium on Electronic Dictionary*. Tokyo, Japan, 1988.
- B. Dorr, G.A. Levow, D. Lin and S. Thomas. Large Scale Construction of Chinese-English Semantic Hierarchy. *Technical Report LAMP TR 040, UMIACS TR 2000-17, CS TR 4120*, University of Maryland, College Park, MD. 2000.
- P. Fung and K. Church. Kvec: A New Approach for Aligning Parallel Texts. In *Proceedings of COLING 1994*, pp 1096—1102, Kyoto, Japan, August 1994.
- P. Fung and Y.Y. Lo. An IR Approach for Translating New Words from Nonparallel, Comparable Texts. In *Proceedings of the 36th Annual Conference of the Association for Computational Linguistics*, pp 414—420, Montreal, Canada, 1998.
- W. Gale, K. Church and D. Yarowsky. A Method for Disambiguating Word Senses in a Large Corpus. In *Computers and the Humanities*, 26:415—439, 1992.
- Y. Hayashi. Translating WordNet Noun Part into Japanese for Cross-Language Natural Language Applications. *Technical Reports of SIG on Natural Language Processing NL 130-10*, pp 73—80, 1999.
- G. Kikui. Resolving Translation Ambiguity using Nonparallel Bilingual Corpora. In *Workshop on Unsupervised Learning in Natural Language Processing*, College Park, MD, 1999.
- K. Knight and S. Luk. Building a Large-Scale Knowledge Base for Machine Translation. In *Proceedings of AAAI '94*, pp 773—778, Seattle, WA, 1994.
- B. Levin. English Verb Classes and Alternations: A Preliminary Investigation. University of Chicago Press, Chicago, IL 1993.
- I.D. Melamed. Automatic Discovery of Non-Compositional Compounds in Parallel Data. In *Proceedings of the 2nd Conference on Empirical Methods in Natural Language Processing*. Providence, RI.
- G.A. Miller, R. Beckwith, C. Fellbaum, D. Gross and K. Miller. WordNet: An On-line Lexical Database. In *International Journal of Lexicography*, 3(4):235—244, 1990.
- M. Palmer and Z. Wu. Verb Semantics for English-Chinese Translation. *Machine Translation*, 10(1-2):59—92, 1995.
- H. Schuetze. Dimensions of Meaning. In *Proceedings of Supercomputing '92*, pp 787—796, Los Alamitos, CA. IEEE Computer Society Press, 1992.
- P. Vossen, P. Diez-Orzas, W. Peters. The Multilingual Design of EuroWordNet. In *Proceedings of the ACL/EACL Workshop on Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*. Madrid, Spain, July 1997.
- P.W. Wong and P. Fung. Nouns in HowNet and WordNet: An Analysis of Semantic Relations. In *Proceedings of the 1st International Conference on Global WordNet*. Mysore, India, 2002.
- D. Yarowsky. Unsupervised Word Sense Disambiguation Rivaling Supervised Methods, In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, pp 189-196, 1995.