Extracting Sense-Disambiguated Example Sentences From Parallel Corpora

Gerard de Melo
Max Planck Institute for Informatics
Saarbrücken, Germany
demelo@mpi-inf.mpg.de

Gerhard Weikum
Max Planck Institute for Informatics
Saarbrücken, Germany
weikum@mpi-inf.mpg.de

Abstract

Example sentences provide an intuitive means of grasping the meaning of a word, and are frequently used to complement conventional word definitions. When a word has multiple meanings, it is useful to have example sentences for specific senses (and hence definitions) of that word rather than indiscriminately lumping all of them together. In this paper, we investigate to what extent such sense-specific example sentences can be extracted from parallel corpora using lexical knowledge bases for multiple languages as a sense index. We use word sense disambiguation heuristics and a cross-lingual measure of semantic similarity to link example sentences to specific word senses. From the sentences found for a given sense, an algorithm then selects a smaller subset that can be presented to end users, taking into account both representativeness and diversity. Preliminary results show that a precision of around 80% can be obtained for a reasonable number of word senses, and that the subset selection yields convincing results.

Keywords

Example Sentence Extraction, Parallel Corpora, Disambiguation, Lexical Databases

1 Introduction

Many dictionaries provide not only definitions but also short sentences that demonstrate how a given word is used in context. Linguists and average dictionary users alike appreciate genuine examples of a word being employed in a sentence.

Goal An example sentence for a word sense is any genuine sentence that contains that word being used in the respective sense. A set of example sentences may (1) allow the user to grasp a word's meaning, and (2) see in what circumstances a word would typically be used in practice.

The first aspect is relevant because traditional intensional word definitions may be too abstract or even confusing to users of a dictionary. Often, the meaning of a word can be determined from its context. In conjunction with conventional definitions, example sentences may allow users to verify whether they have correctly interpreted a definition.

The second aspect is relevant since example sentences may reveal possible contexts a word can be used in. For instance, synonymous words such as '*child*' and '*youngster*' can have the same meaning, yet be used in somewhat different contexts. Examples provide evidence of typical collocations and expressions, e.g. the word '*birth*' often occurs as in '*to give birth*' or '*birth rate*' (but not *'*to give nascence*' or *'*nascence rate*').

For this reason, dictionaries typically include not only conventional definitions, but also example sentences that convey additional information about the meaning of a word. These are often short, limited in number, and in some dictionaries elicited rather than genuine. Hence, retrieving further example sentences can be helpful for lexicographical purposes, or to make the meanings and use more clear to language learners and other laypeople. In modern digital dictionaries, the tight space constraints of print media no longer apply, and thus a larger number of example sentences can be presented to the user on demand.

Our aim is to automatically obtain a set of sense-disambiguated example sentences that are known to mention a specific sense of a word. For instance, for a polysemous word such as '*bat*', we would like to obtain a set of example sentences that refer to the animal sense (e.g. '*There were many bats flying out of the cave.*'), and, separately, a list of example sentences that mention the word in its sports sense (e.g. '*In professional baseball, only wooden bats are permitted.*').

When a user browses a digital dictionary or lexical database, the example sentences could then be provided together with the relevant definitions of the word. Even in digital media, however, more examples may be available than can initially be displayed. For this reason, a means of choosing a restricted set of particularly representative example sentences is an additional requirement.

Contribution Our approach consists of two major building blocks that address the two issues just described. The first step (Section 2) involves extracting the sense-disambiguated example sentences from a parallel corpus by harnessing cross-lingual information to aid in assigning sentences to word senses. The second step (Section 3) selects a limited set of particularly representative example sentences for each word sense, using an algorithm that assesses the contributions made by individual sentences. We provide preliminary experimental results in Section 5.

2 Example Extraction

In the example extraction step, we connect sentences from a corpus to word senses in a given sense inventory whenever we are sufficiently confident that the sentence is an example of the corresponding word being used in the respective sense.

Conventional word sense disambiguation heuristics could be used to determine word senses for a monolingual text, and then the sentences in that text could be linked to the respective senses. Unfortunately, even the most sophisticated all-words disambiguation techniques are currently not reliable enough when a fine-grained sense inventory is used [14].

The intuition behind our method is that, given a parallel text that has been word aligned, we can jointly look at both versions of the text and determine the most likely senses of certain words with significantly greater accuracy than for any single version of the text. After word alignment, we independently apply word sense disambiguation heuristics for each of the languages to obtain ranked lists of senses for each word. One then analyses to what degree the ranked lists for aligned words overlap. In many cases, this makes it possible to infer the sense of a word much more reliably than with conventional disambiguation heuristics. In such a case, we can use the respective sentence in which it occurs as an example sentence for that sense.

Lexical Alignment In the past, parallel corpora had been rather difficult to obtain. This has changed with the increasing multilinguality of the Web as well as the greater demand for such resources resulting from the rise of statistical machine translation. Resnik and Smith [15] showed that the Web can be mined to obtain parallel corpora, while Tiedemann [21] built such corpora from sources such as movie subtitles and manuals of open source software.

To compare the senses of words in both versions of a text, such parallel corpora first need to be word-aligned. This means that occurrences of terms (individual words or possibly lexicalized multi-word expressions) in one language need to be connected to the corresponding occurrences of semantically equivalent terms in the document for the other language.

This is usually accomplished by first aligning sentences, and then using global cooccurrence-based statistics to connect words of two corresponding sentences. Superficial similarities between words and part-of-speech information provide additional clues. We rely on pre-existing tools to perform this alignment, as will be explained in Section 5.

Disambiguation An important prerequisite for our approach is the existence of a word sense database. This resource must provide a fairly complete listing of word senses for a given word in any of the languages involved. We use the WordNet lexical database for the English language and the Spanish WordNet for the Spanish language (see Section 5).

Our system iterates over the sentences in the parallel corpus, simultaneously looking at two different languages a, b. Whenever an occurrence of a word t_a in a is aligned with a word t_b in b, and t_a is believed to be linked to a word sense s_a with a sufficiently high confidence score, we make the sentence where t_a was found an example sentence of s_a.

The confidence score is assigned as follows:

$$\mathrm{score}(t_\mathrm{a}, s_\mathrm{a}) = \mathrm{wsd}(t_\mathrm{a}, s_\mathrm{a}) \frac{\sigma(t_\mathrm{a}, s_\mathrm{a})\mathrm{csim}(t_\mathrm{b}, s_\mathrm{a})}{\sum\limits_{s' \in \sigma(t_\mathrm{a})} \sigma(t_\mathrm{a}, s')\mathrm{csim}(t_\mathrm{b}, s')}$$

The auxiliary function $\sigma(t)$ yields the set of all senses associated with t in the sense inventory, and $\sigma(t, s)$ is the corresponding indicator function ($\sigma(t, s) = 1$ if $s \in \sigma(t)$ and 0 otherwise).

In practice, looking up the possible senses of a word requires a morphological analysis to obtain lemmatized forms of words and determine their part-of-speech. We also rely on a look-ahead window to detect multi-word expressions occurring in the text that have their own sense identifier in the sense knowledge base.

The function $\mathrm{csim}(t_\mathrm{b}, s_\mathrm{a})$ measures the cross-lingual similarity between the likely senses of a term t_b in language b and a specific sense s_a for the word from language a:

$$\mathrm{csim}(t_\mathrm{b}, s_\mathrm{a}) = \sum_{s_\mathrm{b} \in \sigma(t_\mathrm{b})} \mathrm{sim}(s_\mathrm{a}, s_\mathrm{b})\, \mathrm{wsd}(t_\mathrm{b}, s_\mathrm{b}) \qquad (1)$$

These functions build on a monolingual word sense disambiguation function $\mathrm{wsd}(t, s)$ and a sense similarity measure $\mathrm{sim}(s_1, s_2)$.

Monolingual Word Sense Disambiguation The $\mathrm{wsd}(t, s)$ function provides an initial monolingual disambiguation by measuring the similarity between the context of t in the corpus and a similar contextual string created for the sense s. For the former we use the current sentence being disambiguated (which contains t). The latter is created by concatenating glosses and terms associated with the sense s itself or with senses s' directly related via hyponymy, holonymy, derivation, or instance relations, or via up to 2 levels of hypernymy. These context strings are stemmed using the Porter algorithm [13], and feature vectors $\mathbf{v}(t)$, $\mathbf{v}(s)$ with term frequency values are created based on the bag-of-words vector space model. The result is then computed as

$$\mathrm{wsd}(t, s) = \sigma(t, s) \left(\alpha + \frac{\mathbf{v}(s)^T \mathbf{v}(t)}{||\mathbf{v}(s)||\ ||\mathbf{v}(t)||} \right) \qquad (2)$$

Unlike standard word sense disambiguation setups, we prefer obtaining a weighted set of multiple possibly relevant senses rather than just the sense with the highest confidence score. We use α as a smoothing parameter: For higher values of α, the function tends towards a uniform distribution of scores among the relevant senses, i.e. among those with $\sigma(t, s) = 1$.

Semantic Similarity For the semantic similarity measure, we do not rely on generic measures of semantic relatedness often described in the literature [1]. The purpose of this measure here is to identify only word senses that are identical or nearly identical (e.g.

the senses for '*house*' and '*home*') rather than arbitrary forms of association (e.g. between '*house*' and '*door*').

We use the following relatedness measure:

$$\mathrm{sim}(s_1, s_2) = \begin{cases} 1 & s_1 = s_2 \\ 1 & s_1, s_2 \text{ in near-synonymy relationship} \\ 1 & s_1, s_2 \text{ in hypernymy relationship} \\ 1 & s_1, s_2 \text{ in hyponymy relationship} \\ 0 & \text{otherwise} \end{cases}$$

The relational information between senses used here is provided by WordNet.

3 Example Selection

For computational applications, obtaining a repository of perhaps several hundred or even thousand examples for a single word sense can be useful. When displaying examples to human users, it is often better to provide a limited selection at first. The challenge then is deciding which sentences to choose.

We assume there is a space constraint in form of a limit k on the number of sentences that shall be presented to the user. Given a possibly large number of example sentences for a specific word sense, we must choose up to k example sentences that showcase typical contextual collocations and thereby aid the user in discerning the meaning and use of a term.

Assets Each example sentence can be thought of as having certain assets in this respect. For example, for the financial sense of the word '*account*', the fact that an example sentence contains the bigram '*bank account*' could be considered an asset. Another sentence may contain the commonly used expression '*open an account*'.

Our approach looks at 7 different sets of assets (in our case, neighbourhood n-grams) for each example sentence x associated with a word sense.

- $A^1_m(x)$: the original unigram word occurrences for which the example is provided, e.g. '*account*' or '*accounts*' (note that there might be different word forms, and additionally, in WordNet, multiple synonymous words can in fact be associated with a single word sense identifier)
- $A^3_m(x)$: word 3-grams incorporating a preceding and a following word, e.g. '*bank account number*'
- $A^2_p(x)$: word 2-grams incorporating previous words, e.g. '*bank account*'
- $A^3_p(x)$: word 3-grams incorporating previous words, e.g. '*open an account*'
- $A^2_f(x)$: word 2-grams incorporating following words, e.g. '*account manager*'
- $A^3_f(x)$: word 3-grams incorporating following words, e.g. '*account number is*'
- $A^*_m(x)$: the entire sentence

For each of these n-gram sets A^1_m, A^3_m, A^2_p, etc., we also consider the corresponding counter function a^1_m, a^3_m, a^2_p, etc., that counts how often the n-gram occurs in the example sentence in the respective relative position. Usually, this will either be 0 or 1, though an example sentence may also contain multiple occurrences of the word being described, so higher values do occur. Note that in the above use of the words *unigram* and *n-gram*, if the original word being described is a multi-word-expression, it is only counted as one word, e.g. when considering examples for the multi-word expression '*bank account*' instead of just '*account*', the sequence '*opening a bank account*' would be considered a 3-gram.

Our aim will be to choose example sentences that provide representative examples of each of these n-gram sets, so each asset will be given a weight. $A^*_m(x)$, which contains the entire sentence, is a special case where we define $w(a)$ for $a \in A^*_m(x)$ to be the cosine similarity with the gloss context string, as for the word sense disambiguation in Section 2. These weights bias our selection towards example sentences that more clearly reflect the meaning of the word. Apart from this, each n-gram is given a weight based on its relative frequency within the set. For instance, with respect to A^3_p, a frequent expressions like '*open an account*' should receive a much higher weight than '*Peter's chequing account*'. For an n-gram a in the set $A^1_m(a)$, we assign a weight $w(a) = \frac{a^1_m(x,a)}{\sum_{i=1}^{n} a(x_i,a)}$, and equivalently for the other n-gram asset sets $A^3_m(x)$, $A^2_p(x)$, etc.

Objective Of course, at this point one could simply select the top k sentences with respect to the total weight of the n-grams they have as assets. Such an approach however is likely to lead to a very homogeneous result set: n-grams with a high weight occur in many sentences, and hence could easily dominate the ranking.

Instead, we define the goal as follows: Given a set of assets A (in our case, n-grams), a set of items $X = \{x_1, \dots, x_n\}$ (in our case, example sentences), each associated with specific assets $A(x_i) \subseteq A$ (in our case, the union of n-grams returned by A^1_m, A^3_m, A^2_p, etc.), and a limit k, the goal is to choose a set C of items with cardinality $|C| < k$ such that the total weight of the assets

$$\sum_{a \in \bigcup\limits_{x \in C} A(x)} w(a) \tag{3}$$

is maximized.

While this formalization aims at ensuring that items with highly weighted assets occur in the example set, e.g. a sentence containing '*open an account*', it also enforces a certain level of diversity. The latter is achieved by counting the weight of each asset only once, thus if one sentence includes '*open an account*', then there is no direct benefit for including a second sentence with that same n-gram.

The goal can equivalently be expressed in an integer linear program formalization as follows. Define

$$a'(x_i, a) = \begin{cases} 1 & a \in A(x_i) \\ 0 & \text{otherwise.} \end{cases}$$

Our objective is then:

$$\begin{array}{rl} \text{maximize} & \sum_a c_a w(a) \\ \text{s.t.} & c_a \leq c_{x_1} a'(x_1, a) + \cdots + c_{x_n} a'(x_n, a) \\ & c_{x_1} + \cdots + c_{x_n} \leq k \\ & c_a, c_{x_i} \in \{0, 1\} \end{array}$$

This means that we wish to maximize the weight of the assets (n-grams) with $c_a = 1$, where c_a can only be 1 if an appropriate $c_{x_i} = 1$, i.e. an appropriate item (example sentence) x_i is chosen for the result set.

We use a greedy heuristic to find solutions, since the problem is NP-hard.

Proof. We prove the NP-hardness by reducing the NP-hard vertex cover problem to our setting. Given a graph $G = (V, E)$ and a positive integer k, the vertex cover problem consists in determining whether a set of vertices C of size at most k exists, such that each $e \in E$ is incident to at least one $v \in C$. Now set $n = |V|$ and define the items $x_0, \dots, x_n$ to be the vertices $v \in V$. Further, define $A = E$ as the set of assets and $A(x_i)$ as the set of edges incident to x_i. Give these edges uniform weights $w(e) = 1$. Having determined k items that maximize Equation 3, we can then simply test whether the score is equal to $|E|$. If it is, then obviously there exists a set of at most k vertices such that every edge $e \in E$ is covered. If not, then no vertex cover with at most k vertices can exist, because otherwise we could choose that vertex cover as the set of items and obtain a higher objective score (since more edges would be covered). Hence, any vertex cover problem could be answered using an exact algorithm for our problem setting. □

Approach The algorithm we use (Algorithm 3.1) relies on a simple greedy heuristic. It repeatedly chooses the highest-ranked sentence $x \in X$ given the current asset weights w, then resets the weights $w(a)$ of all assets $a \in A(x)$ to zero to ensure that they are no longer considered when choosing further sentences. Ties can be broken arbitrarily (in practice, we first compare the disambiguation scores from Section 2 and choose the highest one).

Algorithm 3.1 Sentence Selection algorithm

```
procedure SELECT($X, k, w$)
    $C \leftarrow \emptyset$
    while $|C| < k \land |X| > 0$ do
        $x \leftarrow \operatorname{argmax}_{x \in X \setminus C} \sum_{a \in A(x)} w(a)$
        $C \leftarrow C \cup \{x\}$
        for all $a \in A(x)$ do
            $w(a) \leftarrow 0$
    return $C$
```

Prior to running the algorithm, an additional filtering may be used. For instance, one may filter out examples that are too long or too short (e.g. incomplete phrases or headlines and titles). One could also allow hiding sentences with possibly offensive or vulgar language.

If the number of example sentences is too large to do a linear scan of all sentences (e.g. in the case of highly frequent words such as conjunctions), we may also choose to let the algorithm run on a smaller random sample $X' \subset X$ of sentences as input.

A useful feature of this greedy algorithm is that it allows emitting a ranked list of entities. Having run the algorithm for a large k, perhaps even $k = \infty$, we can easily obtain the respective output for any $k' < k$ simply by pruning the ranked list generated for k. This can be very useful for interactive user interfaces.

4 Related Work

Several means of generating example sentences for word senses have been proposed. Shinnou et al. [19] extract example sentences for a word from a corpus and attempt to distinguish senses by passing human-labelled sentences as input to a clustering algorithm. This method requires significant human involvement and unlike our approach does not disambiguate senses with respect to a specific sense inventory.

Chklovski and Mihalcea [2] presented a Web interface that asks Web users to tag sentences with the correct word sense and relies on active learning methods to select sentences that are hard to tag automatically.

A different approach suggested by Mihalcea [10] finds example sentences by using a set of seed expressions to create appropriate queries to Web search engines. For example, for the fibre optic channel sense of word '*channel*', appropriate queries would be '*optical fiber channel*', '*channel telephone*', '*transmission channel*'. This method works well when such multi-word constructions can be constructed and could be used to complement our approach.

Another more recent approach [11] clusters words based on a dependency parse of a monolingual corpus. This means that for each word a set of similar words is available. One then tries to match example sentences from the corpus with example sentences already given in WordNet, taking into account the word similarities.

Our approach uses a different strategy by relying on parallel corpora. The intuition that lexical ambiguities in parallel corpora can be resolved more easily has been used by a number of works on word sense disambiguation. Dagan et al. [3] provided an initial linguistic analysis of this hypothesis. Several studies [9, 5, etc.] then implemented this idea in word sense disambiguation algorithms. These approaches are similar to our work. They use simple heuristics on parallel corpora to arrive at sense-labelled data that can then be used *for* word sense disambiguation, while our approach relies on a word sense heuristic to create example sentences from a parallel corpus.

With regards to the challenge of selecting the most valuable examples, Fujii et al. [8] proposed a method for choosing example sentences for word sense disambiguation systems. Unlike our approach, which aims at representative examples for end users, their approach aims at examples likely to be useful for training a disambiguation system. Their proposal selects example sentences that are hard to classify automatically due to the associated uncertainty, so particularly clear examples of a word's use are in fact less likely to get elected. Rychly et al. [17] presented a semi-supervised selection system that learns scores based on combinations of weak classifiers. These classifiers rely on features

Corpus	Covered Senses	Example Sentences	Accuracy (Wilson interval)
OpenSubtitles English-Spanish	13,559	117,078	0.815 ± 0.081
OpenSubtitles Spanish-English	8,833	113,018	0.798 ± 0.090
OpenOffice.org English-Spanish	1,341	13,295	0.803 ± 0.081
OpenOffice.org Spanish-English	932	11,181	0.793 ± 0.087

Table 1: *Number and Accuracy of sense-disambiguated example sentences*

such as word lists, sentence/word length, keyword position, etc. Since the system does not take into account diversity when generating a selection, it would be interesting to combine our algorithm with the scores from their classifiers as additional assets.

5 Results

We conducted preliminary experiments on multiple corpora to evaluate the usefulness of our approach.

5.1 Resources

In terms of parallel corpora, we relied on parts of the OPUS collection [21], in particular the OpenSubtitles [22] and the OpenOffice.org corpora. We made use of GIZA++ [12] and Uplug [20] to produce the word alignments for these corpora. Additionally, we evaluated example sentence selection for undisambiguated sentences using a subset of the Reuters RCV1 corpus [16], consisting of 39,351 documents.

The following lexical knowledge bases were used to build up the sense inventory:

- The original Princeton WordNet 3.0 [7] for the English language.
- The Spanish WordNet jointly developed by three research groups in Spain [6]. Since it was created in alignment with WordNet 1.6, we applied sense mappings [4] to obtain sense identifiers aligned with the version 3.0 of WordNet.

When linking words in the corpus to this inventory, the TreeTagger [18] was used for morphological analysis.

5.2 Experiments

We generated sense-disambiguated example sentences for several setups, and evaluated random samples by assessing whether or not the word was indeed used in the sense determined by our method. The results were generalized using Wilson score intervals, and are presented in Table 1. The smoothing parameter α from Section 2 was set to 0.3. In Table 2, we provide a few anecdotic examples of the output.

In general, this approach yields high-quality example sentences compared to current systems for monolingual text [14]. Automatic word alignment is known to be error-prone, and many heuristics have been proposed to mitigate the effects of this, e.g. aligning in both directions and then intersecting the alignment. In our setting, incorrect alignments are unlikely to lead to incorrect example sentences. This is because two erroneously aligned words in most cases have very different meanings and hence are unlikely to share a semantically similar word sense.

The main cause of the inaccuracies we encountered instead turned out to be the sense inventory's incompleteness. For instance, when an English word has multiple senses shared by the aligned Spanish word, but the sense inventory only lists one of those senses for the Spanish word, our method would lead us to believe that that sense is the right one with high certainty. On a few occasions, incorrect output by the morphological analyser induced errors. For example, when the word '*shed*' was labelled a verb although it was used as a noun, the wrong sense was selected.

A drawback of our approach is that the number of word senses covered is limited. To some degree, this can be addressed by using larger corpora and more language combinations. A reasonably full level of coverage of the senses listed in WordNet would however likely also require relaxing the scoring functions to take into account also less obvious (and hence less reliable) input sentences.

We also applied the sentence selection approach described in Section 3. Table 3 provides ranked lists of example sentences created using Algorithm 3.1. It is clear that frequent collocations such as '*right side*', '*electrical current*', and '*when nightfall comes*' are given a high weight. We also see at least one example sentence wrongly associated with a sense ('*convey*'). Since the algorithm does not depend on sense-disambiguated example sentences, we additionally show sentences from the monolingual RCV1 corpus in Table 4. A larger number of example sentences is typically available here, so the algorithm suceeds even better at choosing sentences that highlight typical collocations, e.g. '*long term*', '*a long time*' for the word '*long*', or '*colonial rule*' and '*colonial power*' for '*colonial*'. The RCV1 corpus is strongly biased towards the financial domain, which is reflected in the example sentences chosen by the algorithm.

6 Conclusions and Future Work

We have presented a framework for extracting sense-disambiguated example sentences from parallel corpora and selecting limited numbers of sentences given space constraints.

In the future, we plan on exploiting alignments with additional languages by using additional versions of WordNet. This would be particularly useful for pairs of languages that are phylogenetically unrelated, as these are more likely to have different patterns of homonymy, and hence a word in one language is less likely to share more than one meaning with a word in the other language.

line (something, as a cord or rope, that is long and thin and flexible)	I got some fishing **line** if you want me to stitch that. Von Sefelt, get the stern **line**.
line (the descendants of one individual)	What **line** of kings do you descend from? My **line** has ended.
catch (catch up with and possibly overtake)	He's got 100 laps to **catch** Beau Brandenburg if he wants to become world champion. They won't **catch** up.
catch (grasp with the mind or develop an understanding of)	I didn't **catch** your name. Sorry, I didn't **catch** it.
talk (exchange thoughts, talk with)	Why don't we have a seat and **talk** it over. Okay I'll **talk** to you but one condition...
talk (use language)	But we'll be listening from the kitchen so **talk** loud. You spit when you **talk**.
opening (a ceremony accompanying the start of some enterprise)	We don't have much time until the **opening** day of Exhibition. What a disaster tomorrow is the **opening** ceremony!
opening (the first performance, as of a theatrical production)	It will be rehearsed in the morning ready for the **opening** tomorrow night. You ready for our big **opening** night?

Table 2: *Samples of Sense-Disambiguated Example Sentences from the OpenSubtitles Corpus (in some cases with multiple words for a single sense identifier)*

The approach could also be extended to *simultaneously* consider aligned sentences from more than two languages to harness example sentences when individual alignments of two languages do not provide enough information for a reliable disambiguation.

For sentence selection, one could consider investigating additional input information for the algorithm, e.g. sentence lengths.

References

[1] A. Budanitsky and G. Hirst. Evaluating wordnet-based measures of lexical semantic relatedness. *Computational Linguistics*, 32(1):13–47, 2006.

[2] T. Chklovski and R. Mihalcea. Building a sense tagged corpus with open mind word expert. In *Proc. ACL 2002 Workshop on Word Sense Disambiguation*, pages 116–122, Morristown, NJ, USA, 2002. Association for Computational Linguistics.

[3] I. Dagan and A. Itai. Two languages are more informative than one. In *Proc. ACL 1991*, pages 130–137, 1991.

[4] J. Daudé, L. Padro, and G. Rigau. Making wordnet mappings robust. In *Proc. 19th Congreso de la Sociedad Española para el Procesamiento del Lenguage Natural (SEPLN)*, Universidad de Alcalá de Henares. Madrid, Spain, 2003.

[5] M. Diab. An unsupervised method for multilingual word sense tagging using parallel corpora: a preliminary investigation. In *Proc. ACL 2000 Workshop on Word Senses and Multilinguality*, pages 1–9, Morristown, NJ, USA, 2000. Association for Computational Linguistics.

[6] J. Farreres, G. Rigau, and H. Rodríguez. Using wordnet for building wordnets. In *Proc. COLING-ACL Workshop on Usage of WordNet in Natural Language Processing Systems*, Montral, Canada, 1998.

[7] C. Fellbaum, editor. *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. The MIT Press, 1998.

[8] A. Fujii, T. Tokunaga, K. Inui, and H. Tanaka. Selective sampling for example-based word sense disambiguation. *Computational Linguistics*, 24(4):573–597, 1998.

[9] W. A. Gale, K. W. Church, and D. Yarowsky. Using bilingual materials to develop word sense disambiguation methods. In *Proc. 4th International Conference on Theoretical and Methodological Issues in Machine Translation: Empiricist vs. Rationalist Methods in MT*, pages 101–112, Montreal, Canada, 1992.

[10] R. Mihalcea. Bootstrapping large sense tagged corpora. In *Proc. 3rd International Conference on Language Resources and Evaluation (LREC), Las Palmas*, 2002.

[11] B. Z. Nik Adilah Hanin and F. Fukumoto. Example-assignment to wordnet thesaurus based on clustering of similar words. *IPSJ SIG Notes*, 2008(46):59–64.

[12] F. J. Och and H. Ney. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51, 2003.

[13] M. F. Porter. An algorithm for suffix stripping. *Program*, 14(3):130–137, 1980.

[14] S. Pradhan, E. Loper, D. Dligach, and M. Palmer. Semeval-2007 task-17: English lexical sample, srl and all words. In *Proc. 4th International Workshop on Semantic Evaluations (SemEval-2007)*, pages 87–92, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

[15] P. Resnik and N. A. Smith. The web as a parallel corpus. *Comput. Linguist.*, 29(3):349–380, 2003.

[16] Reuters. Reuters Corpus, vol. 1: English Language, 1996-08-20 to 1997-08-19, 2000.

[17] P. Rychly, M. Husak, A. Kilgarriff, M. Rundell, and K. McAdam. GDEX: automatically finding good dictionary examples in a corpus. In *Proc. XIII EURALEX International Congress*, Barcelona, Spain, 2008.

[18] H. Schmid. Probabilistic part-of-speech tagging using decision trees. In *Intl. Conference on New Methods in Language Processing*, Manchester, UK, 1994.

[19] H. Shinnou and M. Sasaki. Division of example sentences based on the meaning of a target word using semi-supervised clustering. In *Proc. 6th International Conference on Language Resources and Evaluation (LREC 2008)*, Marrakech, Morocco, 2008.

[20] J. Tiedemann. Combining clues for word alignment. In *Proc. EACL 2003*, pages 339–346, Morristown, NJ, USA, 2003. Association for Computational Linguistics.

[21] J. Tiedemann. The OPUS corpus - parallel & free. In *Proc. 4th International Conference on Language Resources and Evaluation (LREC 2004)*, 2004.

[22] J. Tiedemann. Improved sentence alignment for movie subtitles. In *Proc. RANLP 2007*, 2007.

being or located on or directed toward the side of the body to the east when facing north	1. In America we drive on the **right** side of the road. 2. I'll tie down your **right** arm so you can learn to throw a left. 3. If we wait from the **right** side, we have an advantage there.
put up with something or somebody unpleasant	1. You can't **stand** it can you? 2. You really think I can **tolerate** such an act? 3. No one can **stand** that harmonica all day long.
using or providing or producing or transmitting or operated by electricity	1. Not the **electric** chair. 2. Some **electrical** current circulating through my body. 3. Near as I can tell it's an **electrical** impulse.
take something or somebody with oneself somewhere	1. And they were kind enough to **take** me in here. 2. It **conveys** such a great feeling. 3. We interrupt this program to **bring** you a special news bulletin.
the time of day immediately following sunset	1. When **nightfall** comes go get dressed for the show. 2. You have until **dusk** to give yourselves up. 3. At **dusk** they return loaded with fish.

Table 3: *Example Sentence Rankings (OpenSubtitles Corpus)*

long	1. In the **long** term interest rate market, the yield of the key 182nd 10 year Japanese government bond (JGB) fell to 2.060 percent early on Tuesday, a record low for any benchmark 10-year JGB. 2. "The government and opposition have gambled away the last chance for a **long** time to prove they recognise the country's problems, and that they put the national good above their own power interests", news weekly Der Spiegel said. 3. As **long** as the index keeps hovering between 957 and 995, we will maintain our short term neutral recommendation.
colonial	1. Hong Kong came to the end of 156 years of British **colonial** rule on June 30 and is now an autonomous capitalist region of China, running all its own affairs except defence and diplomacy. 2. The letter was sent in error to the embassy of Portugal – the former **colonial** power in East Timor – and was neither returned nor forwarded to the Indonesian embassy. 3. Sino-British relations hit a snag when former Governor Chris Patten launched electoral reforms in the twilight years of **colonial** rule despite fierce opposition by Beijing.
purchase	1. Romania's State Ownership Fund (FPS), the country's main privatisation body, said on Wednesday it had accepted five bids for the **purchase** of a 50.98 percent stake in the largest local cement maker Romcim. 2. Grand Hotel Group said on Wednesday it has agreed to procure an option to **purchase** the remaining 50 percent of the Grand Hyatt complex in Melbourne from hotel developer and investor Lustig & Moar. 3. The **purchase** price for the business, which had 1996 calendar year sales of about $25 million, was not disclosed.
gold	1. Coach Ian Stacker said his team had hoped to meet the US in the **gold** medal play offs, but because of an early loss to Turkey the team did not get the draw they had counted on. 2. He said India's exports of **gold** and silver jewellery were worth $600 million annually against world trade of about $20 billion. 3. In the bullion market spot **gold** was quoted at $323.80/30 early compared to the London morning fix of $324.05 and the New York close Friday of $324.40/90.

Table 4: *Example Sentence Rankings (RCV1 Corpus)*