The Use of Parallel Corpora in Monolingual Lexicography

How word alignment can identify morphological and semantic relations

Jörg Tiedemann
Department of Linguistics
Uppsala University
Box 527, SE - 751 20 Uppsala, Sweden
Joerg@stp.ling.uu.se

Keywords: morphological relations, semantic relations, word alignment, parallel corpora

Abstract:

Bilingual word alignment can be used to compile lexical data from parallel texts. Word alignment results include a large amount of multiply linked items, which will be the focus of this investigation. Various relations, which are useful for monolingual lexical work, can be seen between translation alternatives that have been identified by the alignment system. In this paper, the nature of these relations is investigated. Several filters are defined in order to categorize them into various kinds of morphological and semantic relations. The filters have been applied to a Swedish/English bitext with promising results as indicated by a small-scale manual evaluation.

1. Introduction

This study investigates the extent to which simple non-linguistic techniques can be used to identify inflectional and other morphological relations, as well as semantic relations such as synonymy and hyponymy from an automatically extracted dictionary. The idea is to take advantage of differences between the two languages and their effect on extracted translation alternatives. Language differences that cause multiple translations
are, e.g., the degree of inflection, the word formation system (e.g. compounding), and the usage of synonyms and homonyms.

Most work related to this study has been done in the field of word sense disambiguation (WSD). In (Dagan et al. 1991), the authors argue that ‘two languages are more informative than one’ when they use parallel text for WSD. Languages differ in the usage of inflectional forms and the expression of certain concepts. The degree of such differences depends on the language pair. In Resnik and Yarowsky (1999), the authors investigate the correlation between language family distance and the tendency to lexicalise differently. They show that WSD is necessary even when translating between closely related languages. Other cross-lingual approaches to WSD are presented in Brown et al. (1991), Gale et al. (1992), and Ide (2000). Possibilities to use statistical data from a corpus in one language for resolving ambiguities in another corpus in a different language are studied in Dagan and Itai (1994). An overview of the state of the art in WSD can be found in Ide and Véronis (1998). Cross-lingual approaches as described above disambiguate multi-sense words by looking at their translations into other languages that might differ with respect to the sense in the context. However, translation alternatives do not have to be due to sense differences. In this study, we focus on relations between translation alternatives that are identified by word alignment.

2. Knowledge-lite Word Alignment

In order to extract translation relations from bitext, we applied a knowledge-lite approach implemented in the Uppsala Word Aligner (UWA). The system comprises several modules for the alignment of words and phrases. It combines statistical
techniques using co-occurrence measures, the identification of cognates through string similarity measures, and iterative processing. Detailed descriptions of UWA can be found in Tiedemann (1999). Evaluation of its performance is presented in Ahrenberg et al. (2000).

3. Techniques for Filtering Dictionaries

Word alignment identifies translation relations in bitext, which can be collected in a dictionary of translation candidates. Only part of the dictionary includes information of value to our investigations. Non-linguistic and knowledge-lite filters can be applied in order to extract candidate pairs with certain relations between them. In this study, the following filters have been applied:

- **String similarity**: Measures such as the Longest Common Subsequence Ratio (LCSR) (Melamed 1995) and other measures based on matching character sequences in two strings (Tiedemann 1997) are used to extract pairs with a certain relation with respect to spelling differences. Similarly, edit distances are also used to compare two strings at the character level.

- **String length and length difference**: A length filter is used to exclude very short strings that often fulfil grammatical rather than lexical functions. The minimal string length was set to four characters. Furthermore, there is often a correlation between the length of tokens and the length of their translations. A length ratio threshold of 0.32 has been used to exclude links with large differences, which is assumed to indicate wrongly aligned items.
- String type filter: Certain tokens are uninteresting for this investigation such as numeric tokens. For this purpose, only purely alphabetic entries have been extracted from the dictionary.

- Frequency based filter: Assuming that links, which have been aligned more frequently, are more likely to be correct than links with a low frequency, the latter can be excluded in order to reduce the amount of wrong alignments.

- Stop-word filter: Grammatical function words usually fulfil several functions with no exact correspondence in other languages. Stop-word lists have been used to remove such links from the base material.

- Stemming: Simple suffix truncation has been applied to source language tokens in order to merge translation alternatives, which were originally aligned to different inflectional variants of the same source language stem.

4. **Extraction of Relation Candidates**

This section will focus on the application of dictionary filters, as described above, in order to extract candidates of different relations from the word alignment dictionary. Dictionary filters can be combined and adjusted such that extracted items show certain characteristics, which are assumed to be specific for the relation under consideration. The following investigation is based on a technical Swedish-English bitext from the MATS corpus with about 100,000 words in each language. The bitext was aligned automatically at sentence and word level. We were mostly interested in relations between Swedish translation alternatives. For this purpose, we swapped the direction of
the extracted dictionary from Swedish-English to English-Swedish, i.e. English terms were linked to one or more Swedish terms.

Swedish and English are quite well suited for this kind of investigation. They differ in their inflectional system and their way of compounding. For instance, Swedish uses different inflectional forms for expressing definite and indefinite forms, Swedish adjectives agree with the nouns they modify, and compositional compounds are frequently used in Swedish compared to non-compositional compounds, which are more common in English.

4.1 Filter 1: Focus on Inflectional Relations

Inflectional variants are usually characterized by close spelling similarities. Inflectional rules are language specific and the degree of inflection varies for different languages. Inflectional changes may include affixation to the common stem as well as minor changes of the common stem (such as ablaut/umlaut). Therefore, we applied a string similarity filter based on LCSR scores (normalization by the length of the longer string) using a rather high threshold of 0.75. Furthermore, inflections usually include only a limited number of modifications. Therefore, we extracted all pairs with edit distances less than five.

Inflectional forms characterize certain word types. The following example shows all inflectional variants of the Swedish compound noun 'felkod' that could be extracted using the filter as described above:

```
fault code     141X: felkod
                120X: felkoder
                72X: felkoden
                52X: felkoderna
                4X: felkodernas
                1X: felkodens
```
The list actually includes all possible inflections of the Swedish noun ‘félkod’ ranked by the linking frequency. Certainly, this is not generally the case. Mostly, all the inflectional forms will not be present in a corpus. Furthermore, certain forms will be overlooked if they do not fulfil the filter constraints. However, many common inflectional forms that have been linked to the same source language stem can be identified. Inflectional patterns have not only been identified for Swedish nouns but also for Swedish adjectives and verbs. In some cases, the filter also extracted derived forms such as compositional forms of particle verbs with a common root, e.g. ‘ipressad’ derived from ‘pressa’. The next filter is focused on the extraction of derived forms.

4.2 Filter 2: Focus on Morphological Relations

We assume that terms that share a common root or stem are related to one another in some sense. Now, we extend our interest from inflectional relations to more complicated types of morphological relations, such as derivation and compounding. Again, string similarity measures can be used to identify common subsequences of two strings.

As in 4.1 we are interested in strings that share a common part, but now we require major differences between two items with a common root. For this experiment, we applied a string similarity measure based on the largest number of character matches at identical positions relative to an offset (Tiedemann 1997) normalised by the length of the shorter string. Using this measure, the filter extracts pairs with similarity scores larger or equal to 0.75. Furthermore, the filter calls for a minimal edit distance of 5.

The following examples illustrate typical outcomes of this filter:
The first example includes both derivations and inflections of the Swedish noun 'last'. The verb 'belasta' is used in its participle form 'belastad'. The definite noun 'belastningen' is derived from the verb 'belasta'. The second example includes another common outcome of the filter, i.e. the compound 'grundjustering', which is a hyponym to the noun 'justering' which has been derived from the verb 'justera' (passive form: 'justeras'). In Swedish as in German, compositional compounds are frequently used. Occasionally they are incompletely aligned to parts of the non-compositional equivalent in English, which is the case in the alignments of the English term 'adjustment'. As a result, an erroneous translation candidate of 'adjustment' appears, i.e. the Swedish compound 'grundjustering'. However, this link can be interesting because it represents a hyponymic relation to the correct translation. Paradoxically, this filtering method gains from weaknesses of the word aligner. Relations of this kind can be observed in many instances in the extracted material.

4.3 Filter 3: Focus on Semantic Relations

We assume that alternative translations, which are not directly derived from common roots, are significantly unlike each other. Using this assumption we can filter out ‘real translation alternatives’ by excluding all translation alternatives which are ‘too similar’ to each other. We do this by applying a string similarity filter with the constraint of a
maximal similarity score of 0.5 and a minimal edit distance of 5. We apply LCSR scores with normalization by the longer string for measuring string similarity.

As mentioned already, word alignment is not perfect. The highest chance to encounter wrongly aligned items probably occurs if focus is set to translation alternatives with low frequencies. Particularly, items that have been aligned only once represent a large proportion of alignment mistakes. Therefore, we decided to exclude all links from the dictionary that have been aligned only once when extracting non-similar translation alternatives.

Assuming correct alignments, there are three possible outcomes of the extraction:

• First, the translation alternates are synonymous to each other in the context in which they appear.

• Secondly, the translated item is ambiguous in the source language and multiple translations exist depending on the sense in the context. The translations, however, are not semantically (or only weakly) related to each other in the target language. Such cases are typically used for word sense disambiguation of source language terms.

• A third category comprises incomplete correspondences, which may include semantically related parts.

Let us consider some examples. The following list presents typical synonyms:

cable       14X:ledningen
             9X:vajer
             4X:kablage
             2X:kabel
reduce      7X:minska
             2X:reducerar
specify     3X:specificerade
             4X:angivet
             2X:angivna
together    14X:ihop
             15X:tillsammans
As illustrated above, synonyms are not restricted to nouns as in the first example but can also be found for verbs (‘minska’ - ‘reducera’), adverbial participles (‘specificerade’ - ‘angivna’), and adverbs (‘ihop’ - ‘tillsammans’).

The following examples illustrate links where the English term is ambiguous whereas Swedish uses distinct expressions for different concepts:

<table>
<thead>
<tr>
<th>term</th>
<th>English</th>
<th>Swedish</th>
</tr>
</thead>
<tbody>
<tr>
<td>work</td>
<td>14X: fungerar</td>
<td>6X: arbeta</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2X: skylt</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3X: plattor</td>
</tr>
<tr>
<td>plate</td>
<td>2X: plattor</td>
<td>13X: anteckna</td>
</tr>
<tr>
<td>note</td>
<td>2X: observera</td>
<td></td>
</tr>
</tbody>
</table>

The English verb ‘work’ can be used in the sense of ‘do a job’ (‘arbeta’), but also in the sense of ‘function’ (‘fungerar’). A ‘plate’ can be a sign with some information (‘skylt’), but also just a ‘plain piece of some material’ (‘plattor’). The last example above presents the ambiguous English verb ‘note’ which can be used in the sense of ‘writing down something’ (‘anteckna’) or in the sense of giving attention to some information (‘observera’).

5. Summary of Results

Non-linguistic and knowledge-lite filters can be combined in several ways. Each parameter can be modified which will change the outcome. Our investigation is concentrated on three configurations, which were presented in the previous sections.

Evaluating the results is not trivial. They should be checked against some reference material from a similar domain. However, such material was not available for this investigation. Manual evaluation is possible, but time-consuming and subjective. However, we carried out a small-scale manual evaluation in order to give an indication
of the success of our method. For this purpose, we extracted 50 random samples from each of the three filtering outcomes and evaluated them by hand.

Table 1 summarizes the number of extracted links and the result of the manual evaluation of each sample.

<table>
<thead>
<tr>
<th>Filter</th>
<th>Inflected</th>
<th>Derived</th>
<th>Compounds</th>
<th>Synonyms</th>
<th>Homonyms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Filter 1</td>
<td>86%</td>
<td>12%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(985 src, 1143 trg, 1758 alt)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Filter 2</td>
<td>16%</td>
<td>74%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(231 src, 295 trg, 623 alt)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Filter 3</td>
<td>14%</td>
<td>28%</td>
<td></td>
<td></td>
<td>16%</td>
</tr>
<tr>
<td>(233 src, 271 trg, 642 alt)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1: Summary of extracted dictionary entries. (src) gives the number of English items. (trg) gives the number of linked Swedish items with a total number of related alternative translations (alt).

The manual evaluation as shown in table 1 gives evidence of the usefulness of the proposed filters. The outcome of the first filter includes mainly inflectional variants. The second filter focuses on morphological relations. Morphological derivations and compounds are included in 90% of the extracted entries. The third filter produces various kinds of semantic relations in about 52% of the cases. The remaining entries contain erroneous and only partly correct links.

6. Conclusion

Knowledge-lite word alignment identifies candidates for translation alternatives from parallel texts. Multiply linked items include valuable information not only about translation relations, but also on morphological and semantic relations between the alternative translation candidates.
We applied simple non-linguistic filters for the extraction of such relations. Clustering characteristics in several categories can be done by simple adjustments of filter parameters. In our approach, we used three different configurations in order to extract inflectional variants, morphological relations, and semantically related items. All the techniques are simple, fast, and adjustable. The results are promising as was indicated in a small-scale evaluation. However, the use of our particular method depends highly on characteristics of the languages under consideration. We have taken advantage of certain differences between English and Swedish. The filters (and the word aligner) have to be modified for investigations of other language pairs.

The proposed approach is meant to support the work of lexicographers and terminologists. Results of such simple processing are useful, not only for augmenting lexical databases with domain-specific material, but also for language checking and the definition of a controlled vocabulary. The extracted material includes valuable hints for the identification of inconsistencies and language errors. Word alignment complemented by filtering techniques produces lexical information with cross-lingual and monolingual relations fast and at low costs. The performance could be significantly improved by using linguistic information. However, non-linguistic approaches produce reasonable results with little effort.

**Bibliography**


