Target Word Selection as Proximity in Semantic Space

Scott McDonald
Centre for Cognitive Science, University of Edinburgh
2 Buccleuch Place, Edinburgh EH8 9LW, Scotland
scottm@cogsci.ed.ac.uk

Abstract

Lexical selection is a significant problem for wide-coverage machine translation: depending on the context, a given source language word can often be translated into different target language words. In this paper I propose a method for target word selection that assumes the appropriate translation is more similar to the translated context than are the alternatives. Similarity of a word to a context is estimated using a proximity measure in corpus-derived "semantic space". The method is evaluated using an English-Spanish parallel corpus of colloquial dialogue.

1 Introduction

When should Spanish *detener* translate to English *arrest* and when to *stop*? This paper explores the problem of lexical selection in machine translation (MT): a given *source* language (SL) word can often be translated into different *target* language (TL) words, depending on the context.

Translation is difficult because the conceptual mapping between languages is generally not one-to-one; *e.g.* Spanish *reloj* maps to both *watch* and *clock*. A SL word might be translatable by more than one TL option, where the choice is based on stylistic or pragmatic rather than semantic criteria. Alternative TL choices also exist for SL words that are ambiguous from the monolingual point of view; *e.g.* English *firm* can be translated by Spanish *firme*, *estricto*, *sólido* or *compañia*.

1.1 Semantic Space Models

In this paper I take a statistical approach to lexical selection, under the working assumption that the *translated* linguistic context can provide sufficient information for choosing the appropriate target. I define the appropriate target as the candidate "closest" in meaning to the local TL context, where *local context* refers to a window of words centered on the "missing" TL item.

To estimate the similarity in meaning between a word and the bag of words forming a context, the semantic properties of words are first represented as their patterns of co-occurrence in a large corpus. Viewing a word as a vector in high dimensional "semantic space" allows distributional similarity (or "semantic distance") to be measured using a standard vector similarity metric. The assumption that distributional similarity corresponds to the psychological concept of semantic relatedness has proved useful in NLP (*e.g.* Schütze, 1992), and for psycholinguistic modelling (*e.g.* Landauer & Dumais, 1997).

One way to estimate the semantic distance between a local discourse context and a target word is to measure the proximity between the centroid vector created from the words in the context and the target word vector. This approach was used successfully by Schütze (1992) in a small-scale word sense disambiguation experiment. However, in this approach the distributional properties of the words making up the local context are not taken into account. The centroid method establishes one position (the mean) on each dimension to use in the distance estimate, without considering the variability of the values on all dimensions. If there is a large amount of noise in the context (semantically irrelevant words), the centroid is influenced equally by these words as by words that are relevant to the correct target. Weighting the dimensions of the space according to variability allows a semantic distance measure to be influenced less by irrelevant dimensions (Kozimo & Ito, 1995).

It is clear that this method relies on the hypothesis that the region of semantic space defined by the translated context "overlaps" to a greater degree with the preferred target than with the alternative choices. The main purpose of the present investigation was to determine the extent that this hypothesis was supported.

1.2 Related Work

Dagan and Itai (1994) have also addressed the lexical selection problem from the TL point of view. Their algorithm uses information about local co-occurrence probabilities for all possible TL pairs of words that can result from translating each pair of words (verb/noun plus argument/modifier) in the SL sentence, and only

makes a decision if the preference is statistically significant. In work aimed at lexical choice in generation, Edmonds (1997) uses information about significant local co-occurrences to choose which of a set of synonyms is most typical in a given context. The present paper differs from these approaches in that *local* co-occurrence behaviour is not considered relevant, but rather an estimate of semantic relatedness between the TL context and each candidate translation.

2 Experiment

To assess the proposed semantic distance (SD) method for target word selection, I used an English-Spanish parallel corpus[1] for testing and evaluation. Several features of a real MT system were incorporated in order that the experiment mimic the type of information available to the lexical selection component. Investigation was restricted to the translation of content words: common nouns, verbs, adjectives and adverbs.

2.1 Materials and Procedure

The test corpus was an English language movie script that had been translated into Spanish on a line-by-line basis. A random sample of 170 lines was extracted from the Spanish half of the corpus, and each content word in this SL subcorpus was looked up in the online version of Langenscheidt's New College English-Spanish Bilingual Dictionary.[2] Experimental items were chosen and a bilingual lexicon (see Figure 1) formed from the information in the dictionary, subject to the following constraints:

- The SL word had two or more *potential* translations.
- A potential translation was defined as a listed translation matching the SL word in POS class (and for verbs, in valency). This simulates the information available from parsing or tagging.
- Only word-to-word translations were considered. Multi-word units in the SL text or listed as a translation were excluded.
- Very low frequency SL words and listed translations (a lexeme frequency of less than 1/million in the 10M word spoken part of the British National Corpus [BNC]) were excluded.

[1]The English half of the corpus consisted of the closed-caption text incorporated with the video release of *Fearless* (Warner Bros/Spring Creek Productions, 1993). The parallel corpus was provided by TCC Communications Corporation, Victoria, BC, Canada.

[2]http://www.gmsmuc.de/english/look.html

```
detener ⇒ stop arrest detain delay hold
mejorar ⇒ improve increase
precio  ⇒ price cost value worth
```

Figure 1. Example bilingual lexical entries.

The translations given in the parallel corpus for 13 SL items were not listed in Langenscheidt's. This was due to the directionality of bilingual dictionaries – entries are created from the TL point of view – and the fact that the direction of original translation was opposite to that used for building the testing lexicon. These translations were incorporated into the bilingual lexicon. A total of 99 experimental items were compiled.

For each SL item, the corresponding TL translation was located in the parallel corpus and all TL content words within a ±25 word window were extracted to form the local discourse context. Co-occurrence vectors for each lemmatised context word meeting the frequency threshold were created from a lemmatised version of the spoken part of the BNC. Vectors were constructed by advancing a window of ±3 words through the corpus, and for each word recording the number of times each of 446 index words occurred within the window. This procedure produced a 446-dimension semantic space. Finally, co-occurrence counts were replaced with their log-likelihood values, which effectively normalizes the vectors. Parameter settings were taken from McDonald (1997). Vectors for the translation candidates were created using exactly the same method.

Compared to a practical MT system, the lexical selection simulation makes several simplifying assumptions. For one, two or more items in the same SL sentence are treated as if all other items are already correctly translated. Secondly, the use of forward context means that a word is left untranslated until a prespecified number of following words are translated. Finally, the bilingual lexicon listed 4.2 translation candidates per entry on average. Many of the alternatives could be described as stylistic variants, and might not be present in an actual MT lexicon.

2.2 Calculating Semantic Distance

The proximity of each translation candidate to the bag of words forming the local TL context was measured as described below, and the "closest" target was chosen. The method for scaling each dimension of the space was adapted from Kozimo and Ito (1995) in order to de-emphasize dimensions irrelevant to the local

context. If the variability of vector component i is high, then this dimension is considered to be less relevant than a component with lower variability, and the semantic distance measure should take this into account.

The relevance r_i for each dimension is defined as the ratio of the standard deviation s_i of the distribution formed by dimension i, for all local context words LC, over the maximum standard deviation s_{max} for LC:

$$r_i = \frac{s_i}{s_{\max}}$$

For each candidate translation t the vector representing each word c in LC is moved to a new position in the space according to a function of r and its current distance from t:

$$c_i' = c_i + r_i(t_i - c_i)$$

If r is large, then any difference in the value of component i between t and LC is made less prominent than if r is small. Finally, semantic distance is calculated as the mean cosine of the angle between target t and each word c in LC:

$$SD(t, LC) = \frac{1}{|LC|} \sum_{c' \in LC} \cos(t, c')$$

2.3 Results and Discussion

Performance was evaluated against the actual English translation aligned with each Spanish item. Two baseline measures were used for comparison: accuracy expected by random selection, and word frequency (WF; selection of the translation candidate with the highest corpus frequency). The semantic distance method made 57/99 correct choices (57.6%) whereas the frequency method bettered it slightly, choosing the aligned translation 59 times (59.6%). Expected chance performance was 22.9%. Of the errors made by WF, SD corrected 15%, and WF corrected 19% of the SD method's errors.

In about one-quarter of the errors made by the SD method, the selected candidate and the "correct" translation seemed equally acceptable in the context. This can be seen more clearly in an example TL context for *trabajo* (Figure 2). There appears to be little information available in the context in order to prefer *work* over the closely related *job*.

Performance was assessed at the level of 100% applicability – the SD method was used for *every* item. Future work will investigate the use of a confidence estimate: if the evidence for preferring one candidate over another is weak, an alternative selection method should be used.

```
SL: Ud. es muy dedicado a su trabajo.
TL: ... to go back to the office.
    what's your name?
    i'm john wilkenson.
    why were you on the plane?
    on business.
    you're very committed to your <X>.
    you go ahead and finish your story,
    please.
    we were taking a vacation--
    my sister, me, and our kids.
    you know--
    no husbands.
    we saw ...
```

Figure 2. Example discourse context for alignment *trabajo⇒work*. X indicates the target word position.

3 Conclusion

A preliminary investigation of a method for lexical selection in MT was presented. The assumption that the preferred translation of a translationally ambiguous SL word is the one closest in semantic distance to its translated context gave encouraging results, taking into account the impoverished nature of the information available in spoken language context.

Acknowledgements

This work was supported by awards from NSERC Canada and the ORS scheme, and in part by ESRC grant #R000237419. Thanks to Chris Brew and Mirella Lapata for valuable comments.

References

Dagan, I. & A. Itai. 1994. Word sense disambiguation using a second language monolingual corpus. *Computational Linguistics*, 20:563-596.

Edmonds, P. 1997. Choosing the word most typical in context using a lexical co-occurrence network. In *Proceedings of the 35th ACL/8th EACL*, Madrid.

Kozima, H. & A. Ito. 1995. Context-sensitive measurement of word distance by adaptive scaling of a semantic space. In *Proceedings of RANLP-95*, pages 161-168, Tzigov Chark, Bulgaria.

Landauer, T. K. & S. T. Dumais. 1997. A solution to Plato's problem: the Latent Semantic Analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104:211-240.

McDonald, S. 1997. Exploring the validity of corpus-derived measures of semantic similarity. Paper presented at the *9th Annual CCS/HCRC Postgraduate Conference*, University of Edinburgh.

Schütze, H. 1992. Dimensions of meaning. In *Proceedings of Supercomputing '92*, pages 787-796, New York: Association for Computing Machinery.