Dreams of a Universal Translator: how language-independent can statistical machine translation be?

Joshua Herring
Department of Linguistics
Indiana University
jwherrin@indiana.edu

Abstract

This paper follows on work by (Piao, 2002), which surveyed the efficacy of a number of statistical word-alignment correlation metrics over an English-Chinese corpus, repeating the method for the Vermobil German-English aligned corpus. For the most part, it was found that the same metrics ranked high on efficacy in both surveys, indicating that these methods are probably independently reliable regardless of language pair.

Overview

Statistical Machine Translation (SMT) is a highly practical field - often more concenred with getting good results than making scientific generalizations. This is of necessity: natural languages vary in their properties, and features of the surface string that are of high linguistic significance in one language (say, the leading and ending characters of words in morphologically rich languages like Finnish) may simply not exist in another (an isolating language like Vietnamese, for example). To the extent that success depends on tailoring the system to exploit idiosyncratic properties of the source and target languages, language-independent systems seem like a pipe dream. But what if the process of tailoring the system could itself be automatic?

This question is strongly hinted at in (Piao, 2002), which proposes a hybrid algorithm for statistical parallel-corpus word alignment for Chinese and English. Much of the motivation in that work comes from the fact that earlier statistical alignment mechanisms made assumptions which are inappropriate for this language pair. For example, (McEnery and Oakes, 1996) make use of morphological variants in aligning English and French texts, a method that only works for pairs of languages which employ reasonably similar morphological systems. Likewise, (Melamed, 1995) reports that POS tagging helps in word alignment. This information again comes, however, from a French-English pairing. Semantically similar items in Chinese are not as likely to share a common grammatical class with their English counterparts, and methods which rely to any significant extent on POS similarity cannot be expected to perform as reported for Chinese. Thus, the focus was on developing a statistical method that can be relied upon regardless of language pair. To this end, (Piao, 2002) surveys a number of suggested alignment similarity metrics from the literature and tests them for efficacy on a corpus of sentence-aligned pamphlets written in English and transated into Chinese.

Purpose

As (Piao, 2002) takes as its aim the creation of an alignment method which does not rely on arbitrary similarities in the idiosyncratic properties of the langauge pair under operation, it is worth asking whether the method presented there would get similar results on language pairs other than English-Chinese. The purpose of this project was to repeat this experiment for a German-English corpus aligned at the sentence level and compare the results with what Piao found for Chinese. There was no expectation for how individual metrics would perform; the work here was purely exploratory.

Background - Piao 2002

The task of word alignment is to identify true word translations contained in bilingual or multilingual parallel corpora. Automatic statistical word alignment methods tend to be based on the heuristic that translational equivalents will have similar (or at least highly correlated) positional distributions across each half of a parallel corpus. (Piao, 2002) begins with a review of the most common statistical metrics used to determine positional similarity, critiques some of the assumptions of previous approaches as unsuitable for English-Chinese parallel corpora, and then proposes (and tests) a hybrid algorithm that is meant to address some of the issues raised.

In general, two types of metics are considered: contingency table metics and pure positional metrics – called *dispersion* metrics.

Contingency Table Metrics draw up a collection of correlational measures between two candidate items, one from each of the two languages in the pair under study. In general, it is assumed that the corpus has been aligned by sentence - or else by simple arbitrary separation of the text into a predetermined number of (roughly) equally-sized "chunks." A table of values is then drawn up based on the occurrence or lack thereof of each word in the sections. In (Piao, 2002), these are indeed aligned sentences:

	x	
y	a	b
	c	d

Where:

y is a word in the source language

x is a word in the target language

a is the number of sections in which the two words both appear

b is the number of sections in which x appears without y

c is the number of sections in which y appears without x

d is the number of sections which contain neither x nor y

It is assumed that *all* words which appear in the target text (save a handful of enumerated "stop" words which contain distributions likely to confuse the metric) are potential translations for a word in the source text. Consequently, contingency tables like the one above are drawn up for all such potential pairs. Cooccurence significance metrics are then used to identify those pairs which seem to be significantly correlated. Pairs which meet a predefined threshold are considered likely to represent actual translational equivalents.

Dispersion Metrics One problem noted with contingency table metrics is that individual items often give rise to multiple translations, depending on context. Thus, metrics based on co-occurence may in some cases fail to detect a significant correlation because of polysemy in the source word, or else because the meaning of the source word has multiple expressions in the target language. A suggestion for correcting for this comes in the form of *word distribution distance measures* - called *dispersion metrics* here. The idea behind a dispersion metric is that words of similar meaning will congregate around similar areas of the text - even if they are not necessarily direct translations of each other in all cases(Fung and Church, 1994).

Dispersion metrics operate by dividing the source and target texts into proportional "chunks" and noting the frequency of each item in each "chunk." A vector of token frequencies across chunks is formed for each item - e.g. `< 1, 3, 1, 0, 0>` for an item that appeared once in the first section, thrice in the second section, not at all in the last section, etc. Such vectors then feed the calculation of a general term that captures the relative "evenness" of the item's distribution across the text, with high values typically indicating a relatively balanced distribution, and low values indicating that an item clusters around certain sections only. As the resulting term captures only the degree to which an item's distribution across the text is skewed or balanced and not in which sections a term with a skewed distribution tends to appear, such measures are genenrally concerned with generating a lexicon of potential "anchors" which can be used to feed later stages of a multi-stage alignment algorithm rather than being employed for direct alignment themselves. In (Piao, 2002), they see use mostly as filters to resolve lingering ambiguities from the application of the contingency table metrics.

The algorithm

The "hybrid algorithm" proposed in (Piao, 2002) is actually a four-pass algorithm for lexicon creation. The first pass creates a list of "stop words" - items that occur often enough to cause spurious statistical alignments (those with a Julliand D score greater than 100 in the work being surveyed) - and removes them. The second pass attempts to find alignments within pre-aligned sentence pairs based on a range of contingency-table-based cooccurrence measures. The third pass pares down the list of alignments so hypothesized by removing any that do not meet significance according to a number of dispersion measures. If any ambiguities remain, a fourth pass removes them with a POS filter.

What is suggestive about Piao's approach is that he tested several combinations of well-known statistical measures at each stage of the algorithm - pairing various cooccurrence metrics in step two with various dispersion metrics in step three, etc. This raises the possibility of systems that, given a list of strategically-chosen translation pairs, could determine for themselves which (combinations of) metrics were useful for an arbitrary input corpus.

Method

For the purpose of testing Piao's results, a software tool - LINEMAN - was developed to allow for rapid selection of pairs of statistical metrics. It is essentially a GUI interface that allows the user to select from, order, and set significance levels for the statistical methods surveyed in the paper, as well as select input texts and post-run evaluation statistics. Using this tool, the tests detailed in Piao were run on the German-English section of the Vermobil corpus and compared to the results Piao obtained. Cooccurrence measures surveyed were the Phi Squared (Φ^2) measure of (Gale and Church, 1991), various Mutual Information scores, and all the metrics used in (Daille et al., 1994) (the Simple Matching Coefficient, the Kulczinsky Coefficient, the Ochiai Coefficient, Fager and McGowan Coefficient, the Yule Coefficient, and log-likelihood ratio). Dispersion measures used included Caroll's D, Julliand's D, and Rosengren's S (see (Lyne, 1985) for a description). One difference from Piao's work, however, is that no POS filter was used here - that is, the final pass of the algorithm was cut out[1]. Formulae for the various metrics used are given in the appendix.

Step One: Stop Words

Unfortunately, (Piao, 2002) goes into very little detail about the choice of stop words, saying only that a list of all words with Julliand's D scores higher than 100 was generated and then "hand-adjusted." It is unclear what guided this adjustment, and no list of sample words added or removed from the original list was included. However, as Piao's corpus consisted of translations of pamphlets, it is probable that certain words were highly frequent in the source text that would not be in a more balanced corpus. It is therefore probably reasonable to assume that what is meant by "hand-adjusted" is that certain key terms were removed from the list of automatically-identified stop words on the grounds that they were likely to be effective anchors. It may also have been the case that some care was taken to insure that the number of items so removed was roughly the same for each language pair. As these are issues less likely to have been present for the Vermobil corpus, which is reasonably representative of German and English speech, and as there was in any case no way to know for sure what "hand-adjustment" methods Piao employed, the present study elected to throw out all words which appeared on the automatically-generated list.

Step Two: Contingency Metrics

After all words on the stop list were removed, the cartesian product of the set of remaining words for German and the set of remaining words for English was taken to generate a list of potential alignment pairs. A contingency table was drawn up for each pair, and a score obtained for each of the metrics under consideration.

Piao reports that significance thresholds for each metric were "obtained empricially," but the decisions reached are not included in the published version of his survey. Consequently, this project simply retained the pairs which received the 5 highest scores for each metric. In cases where multiple pairs received the same score, all such pairs were retained, resulting in lists longer than 5 in a handful of instances. The choice of 5 was arbitrary here, designed mostly to replicate the number that Piao seemed to be obtaining for his results while keeping the amount that had to be handled manageable.

Step Three: Dispersion Metrics

Piao's dispersion metric filter operates according to Euclidean distance over vectors formed from scores from some combination of the metrics. As all possible such combinations were used in Piao (including vectors of only one, and a vector formed from all three), this study also generated scores for all possible combinations for each wordpair. Each list generated in the previous step was then filtered according to the threshold scores obtained for dispersion metrics reported in (Piao, 2002). Remaining candidates formed the output of the process.

It should be noted that Piao included an additional filter – called *Word Frequency Distance* – formed from the Euclidean distance between vectors formed form the relative frequencies of each item in a potential alignment pair by subsection. That is, for each "chunk" n, there was corresponding term in the vector representing the candidate item's relative frequency in that section. The WFD score is the simple Euclidean Distance between the vectors so formed for candidate items in each half of the corpus. Because

[1] This was due to time contraints. It would, of course, be interesting to see how much a POS filter helps, but Piao was not specific about the operation of his, and coding one specifically for LINEMAN seemed a distraction from the main point of the exercise, which was the comparison between the statistical measures

Piao reports somewhat idiosyncratic results for this filter (it has high precision but relatively low recall), it was not included in the initial survey here due to time constraints, despite a generally favorable showing in (Piao, 2002). It will be included in future versions of the project.

POS Filter

As Piao gives no infomration about the functioning of the POS filter he employed for the final step in his algorithm, and as the use of this filter was anyway intended to be a last resort for resovling any ambiguities that remained after step three, no such filter was included in this project.

Overview of Results

There was significant overlap with Piao's results, indicating that his findings are language-independent - at least to a certain degree. In particular, this survey confirmed that the MI^2 method - a mutual information measure - fared best among the cooccurrence methods, outperforming even the MI^3 and log-likelihood ratios. Likewise, for the dispersion metrics it was found that a combination of Caroll's D and Julliand's D fared best, and that Rosengren's S was not as useful. Best of all was a combination of all three metrics. Interestingly, however, all results are slightly less robust than for Piao. That is, the same *pattern* of results was obtained, but the overall level of success was lower. This probably owes to two factors. First, (Piao, 2002) is not entirely forthcoming about the method behind the creation of the "stop list," but it was clearly more selective than the method employed in this survey. It is possible that the automatically-generated "stop list" in this survey threw out some salient alignment pairs. For similar reasons, the process in step two was markedly different from that used in (Piao, 2002). It is uncertain how this difference affected the outcome, but as no significance tests were done in step two here it is probable that this survey retained more spurious alignment pairs than Piao's survey.

Rankings of metrics here were scored using the same E-score for combining precision and recall reported in (Piao, 2002). A complete table of results can be downloaded from `http://mypage.iu.edu/~jwherrin/mclc08/results.pdf`.

Discussion and Future Directions

In some ways, the frustrations in trying to reproduce Piao's methods exactly make the similarities between the patterns of results seem all the more robust. Though there were almost certainly non-trivial differences in the approaches taken, the same metrics show themselves to be most useful for both corpora. This is encouraging for any researchers who want to establish empirically that certain of these methods work better regardless of language pair, and especially for the possibility of choosing among them for a "universal" language-independent system.

More encouraging for the possibility of developing a language-independent SMT system, however, is the fact that similar results were obtained despite the differences in method for Step Two. Piao reports, without elaboration, that thresholds were adjusted "empirically." No such thresholds were used in this survey, *prima facie* evidence that they may not be crucial. If this proves to be true for more corpora formed from different langauge pairs, it could seriously simplify the design of a language-neutral system.

Obviously there is a lot of potential for followup research. Most pressing is the need to replicate this experiment for still more language pairs. Also crucial, however, is the need to obtain more precise results for how adjustments to the various parameters – especially the sizes of the "chunks" used in the dispersion metrics – affect the overall results.

APPENDIX: Metrics Used

Contingency Metrics

Variables here are as outlined in previous sections.

1. Simple Matching Coefficient

$$SMC = \frac{a+b}{a+b+c+d}$$

2. Kulczinsky Coefficient

$$KUC = \frac{a}{2}\left(\frac{1}{a+b} + \frac{1}{a+c}\right)$$

3. Ochiai Coefficient

$$OCH = \frac{1}{\sqrt{(a+b)(a+c)}}$$

4. Fager and McGowan Coefficient

$$FAG = \frac{1}{\sqrt{(a+b)(a+c)}} - \frac{1}{2\sqrt{(a+b)}}$$

5. Yule Coefficient

$$YUL = \frac{ad-bc}{ad+bc}$$

6. McConnoughy Coefficient

$$\frac{a^2 - bc}{(a+b)(a+c)}$$

7. Phi-square Coefficient

$$\Phi^2 = \frac{(ad-bc)^2}{(a+b)(a+c)(b+c)(b+d)}$$

8. Association Ratio (MI)

$$MI = \log_2 \frac{a}{(a+b)(a+c)}$$

9. Squared Association Ratio (MI^2)

$$MI^2 = \log_2 \frac{a^2}{(a+b)(a+c)}$$

10. Cubic Association Ratio (MI^3)

$$MI^3 = \log_2 \frac{a^3}{(a+b)(a+c)}$$

11. Log Likelihood

$$\begin{aligned} \mathit{LogLikelihood} &= a\log a - b\log b - c\log c - d\log d \\ &-(a+b)\log(a+b) - (a+c)\log(a+c) \\ &-(b+d)\log(b+d) - (c+d)\log(c+d) \\ &+(a+b+c+d)\log(a+b+c+d) \end{aligned}$$

Dispersion Metrics

1. Julliand's D

$$D = 1 - \frac{V}{\sqrt{n-1}}$$
$$V = \frac{s}{\bar{x}}$$

2. Caroll's D_2

$$D_2 = \frac{H}{\log_2 n}$$
$$H = \log_2 P - \frac{\sum p \log p}{P}$$

3. Rosengren's S

$$S = \frac{KF}{F}$$
$$F = \sum x$$
$$K = \frac{1}{n}\left(\sum \sqrt{x}\right)^2$$

References

Daille, Beatrice, Eric Gaussier and Jean-Marc Lange (1994). Towards automatic extraction of monolingual and bilingual terminology. In *Proceedings of the 15th conference on Computational linguistics - Volume 1*. Association for Computational Linguistics, pp. 515 – 521.

Fung, P. and K. Church (1994). Kvec: A new approach for aligning parallel texts.

Gale, William A. and Kenneth Ward Church (1991). A Program for Aligning Sentences in Bilingual Corpora. In *Meeting of the Association for Computational Linguistics*. pp. 177–184.

Lyne, A. A. (1985). Dispersion. In *The Vocabulary of French Business Correspondence*, Slatkine-Champion.

McEnery, A. and M. Oakes (1996). Sentence and Word Alignment in the CRATER Project. In J. Thomas and M. Short (eds.), *Using Corpora for Language Research*, Longman, pp. 211 – 31.

Melamed, D. (1995). Automatic evaluation and uniform filter cascades for inducing n-best translation lexicons. In Christian Boitet and Pete Whitelock (eds.), *Proceedings of the Third Workshop in Very Large Corpora*. Boston, MA, pp. 184–98.

Piao, Scott Songlin (2002). Word Alignment in English-Chinese Parallel Corpora. *Literary and Linguistic Computing* 17(2).