Predicting alignment performance

Lars Ahrenberg

Department of Computer and Information Science
Linköping University
lars.ahrenberg@liu.se

Abstract

We present a planned project aimed at evaluating the performance of different word aligners under different conditions. We want to be able to choose the best word aligner for a given corpus and purpose, and, more specifically, to predict a word aligner's performance on the basis of properties of the data. Some of these properties relate to the languages used, while others concern the nature of the parallelism. Performance will be measured wrt both token matches and matches of translation units.

1. Background

It is not uncommon that the output of a system based on statistical learning, such as Google Translate or Giza++ makes you disappointed. This is often due to a mismatch between the data given to the system and its models. Even for a researcher, the success or failure of a given task can be hard to predict. In the case of word alignment there is a rich smorgasbord of systems to choose from, including Giza++ (Och and Ney, 2003), Berkeley Aligner (Liang et al., 2006), Uplug clue aligner (Tiedemann, 2003) and more. Furthermore, each system has a number of parameters that can be set by the user, making the decision even more complex.

2. Goals

The overall goal of the project is to get a better understanding of the adequacy of different word aligners for different types of alignment tasks. In particular, we want answers to questions such as the following:

Given a corpus that we wish to word align, which aligner should we choose? First, the purpose of the alignment is of course an important factor. If the purpose is statistical machine translation, the resulting word alignment will be used by another process that creates data for the decoder, and a common understanding is that we should then go for high recall. However, if the purpose is to build a word aligned resource where translation phenomena can be looked up, high precision should not be sacrificed for better recall. Another relevant aspect is the availability of resources. With millions of aligned sentence pairs available, it is a good choice to use word aligners that employ statistical models, but if the corpus is limited to a few thousand sentence pairs, these may not produce satisfactory results. Similarly, if we have a dictionary, we should quite obviously look for a system that can make good use of it.

A third aspect, which is the one in focus in the project, is the *corpus features*. There are a number of features that are known to affect word alignment systems negatively, such as differences in morphological complexity of the languages concerned, the occurrence of non-local reorderings and null matches. But exactly how such factors affect the outcome is less known. This brings us to the next question:

How can we explain the performance of a word aligner on a given parallel corpus? The general answer to this question is to be found in the fit (or lack thereof) between the features of the corpus and the alignment models used by the system. To give a more useful answer, we need to provide a detailed account of the corpus features and relate them to the system models. I call such a detailed account of corpus features an *alignment profile*. An alignment profile can be defined as a set of probability mass functions that show how token matches are distributed over types. See Table 1 for some examples.

If we know the constraints on alignments that a system assumes, we can be definite about what it can not find. But we cannot know that it will find all instances of what it is looking for. This calls for more empirical studies and brings us to a third question:

How well can we predict the performance of a word aligner, as measured by precision, recall or error rate, from an alignment profile of the test data? (Birch et al., 2008) studied the effects of differences in linguistic structure on machine translation performance, using three dimensions of difference: reordering, morphological complexity, and language relatedness. They concluded that each of these three factors has a significant effect on translation performance as measured by the BLEU score, and a combined model could account for 75% of the variability of the performance for 110 language pairs. They did not report figures on alignment per se, however. But, arguably, word alignment performance should be possible to predict equally well.

A problem, though, is that there exist more reference data for translation than for word alignment. To handle that problem we must use artifical data.

Type description	Examples
Number of tokens	0-1, 1-1, 2-1, m-m, ...
Token positions	same, close, far, ...
Corpus frequency	1, 2, 6-9, ... 100+

Table 1: Dimensions of typing for translation units.

3. Method

We want to test different word aligners with data that vary in alignment profiles. Given an inventory of primitive types, as in Table 1, we can go on to define complex, descriptive properties of alignments in terms of the basic types. For example, we may define *neatness* as the percentage of units that are 1-1, and *faithfulness* as the percentage of all tokens that have received a non-null match.

Language-related properties such as differences in morphological complexity and lexical similarity can also be studied. (Birch et al., 2008) found that differences in type-token ratios, a metric that reflects both these causes, accounted for about a third of the variation in translation performance as measured byn BLEU.

3.1 Metrics

Research on word alignment has mostly been performed in the context of statistical machine translation (SMT) and been evaluated on the basis of machine translation performance. (Och and Ney, 2003) also evaluated intrinsically using Alignment Error Rate (AER) as their measure. This metric has been criticized for good reasons, and with our goals in mind, the major drawback is that it is too coarse and does not reveal qualitative differences. Other common metrics are precision and recall, usually measured on the set of token matches (or links). (Søgaard and Kuhn, 2009) defined a measure they called Translation Unit Error (TUER) which assumes that the alignment is complete and unit-based. This means that there is a decision for all tokens, conforming to the constraint that if tokens i, i', j, j', if $< i, j >$, $< i, j' >$, and $< i', j >$ are aligned, then so is $< i', j' >$. Metrics based on translation units are actually more relevant for purposes of resource creation and will be used in the project.

3.2 Corpus generation

Available natural gold standards can be used, where available, but to systematically study the effects of different alignment profiles, we need to be able to generate data with known properties. For this purpose we use probabilistic synchronous grammars.

The synchronous grammars generate sentence pairs with their word alignments. Null alignments as well as many-to-many alignments (including many-to-one and one-to-many) can be generated and the frequency of these alignments is determined by the probabilities assigned to rules that define them. Similarly, the amount of reorderings in the corpus is determined by the probabilities of the rules that have reordered constituents. Some rule examples are illustrated in Figure 1.

The vocabulary is divided into parts-of-speech with a different treatment of function words and content words. Each content word, for both source and target vocabularies, is associated with one or more *cepts* where a cept represents a meaning. The cepts determine possible alignments. Multiword expressions start life as single tokens in the grammar and are then split in a separate process to produce many-to-many alignments.

The current grammars have rules of depth one and are thus not expressive enough to be able to generate all types of alignment phenomena that occur in real translations (Søgaard and Kuhn, 2009). Still, the types of alignments they can generate allow for a wide range of alignment profiles.

NP → N, N, 1-1, 0.46
NP → N, P N, 0-1 1-2, 0.10
NP → AP N, N AP, 1-2, 2-1, 0.10
A: 500, 100, 50, 0.06, 0.04
N: 4000, 500, 100, 0.05, 0.04
P: 20, 20, 10, 0.16, 0.12

Figure 1: Examples of rules and vocabulary definitions. Positive numbers indicate the number of lexical items to be generated with one, two or three meanings. Numbers between 0 and 1 are probabilities.

Scores	**Gold**	**IBM-1**	**IBM-2**	**IBM-4**
Precision	1	0.843	0.888	0.929
Recall	1	0.785	0.831	0.858
Faithfulness	0.963	0.916	0.949	0.929
Neatness	0.807	0.780	0.840	0.886

Table 2: Scores and profiles for three alignment models compared with a gold standard.

In Table 2 we show data from a run of Giza++ on an artificial corpus with 50,000 sentence pairs. Sentences varied in length between 2 and 100 tokens with an average of 10.4. The vocabulary defined by the grammar had just under 7000 source stems and some 8400 target stems. The morphology was simple with an equal number of inflections for both languages. As is expected precision and recall improve with better models, but all systems' alignment profiles differ from the that of the gold standard. In particular, we can see that IBM-4, which has by far the best scores in terms of precision and recall, exaggerates the neatness and underestimates the faithfulness of this corpus.

4. References

Alexandra Birch, Miles Osborne, and Philipp Koehn. 2008. Predicting success in machine translation. In *Proceedings on Empirical Methods in Natural Language Processing (EMNLP)*, pages 745–754, Honolulu, USA, October.

Percy Liang, Ben Taskar, and Dan Klein. 2006. Alignment by agreement. In *Proceedings of HLT-NAACL*, pages 104–111, New York City, USA, June.

Franz J. Och and Hermann Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1):19–51.

Anders Søgaard and Jonas Kuhn. 2009. Empirical lower bounds on alignment error rates in syntax-based machine translation. In *Proceedings of SSST-3 Third Workshop on Syntax and Structure in Statistical Translation*, pages 19–27, Boulder, Colorado, June.

Jörg Tiedemann. 2003. Combining clues for word alignment. In *Proceedings of the Tenth Conference of the EACL*, pages 339–346.