Bilingual Lexicon Construction Using Large Corpora

CS TR 3666, UMIACS TR 96-50

Wade Shen
Bonnie J. Dorr
Department of Computer Science
University of Maryland College Park
{swade,bonnie}@cs.umd.edu

June 26, 1997

Abstract

This paper introduces a method for learning bilingual term and sentence level alignments for the purpose of building bilingual lexicons. Combining statistical techniques with linguistic knowledge, a general algorithm is developed for learning term and sentence alignments from large bilingual corpora with high accuracy. This is achieved through the use of filtered linguistic feedback between term and sentence alignment processes. An implementation of this algorithm, TAG-ALIGN, is evaluated against approaches similar to [Brown et al.1993] that apply Bayesian techniques for term alignment, and [Gale and Church 1991] a dynamic programming method for aligning sentences. The ultimate goal is to produce large bilingual lexicons with a high degree of accuracy from potentially noisy corpora.

Introduction

Given that bilingual text corpora have become widely available in electronic form, researchers have sought methods for mining the cross-language information they embody. Much of this research has focused on techniques to align sections of documents (typically sentences) with their respective translations. Such techniques employ statistical measures to find best-fit mapping between document sections and their corresponding sections in a translated document.

This paper discusses a system, TAG-ALIGN for producing sentence and word level alignments. The algorithm, in its alignment process, generates a highly accurate bilingual lexicon that can be used to generate better word and sentence level alignments from new corpora.

TAG-ALIGN's word alignment algorithm draws on a Bayesian model similar to that discussed in [Brown et al.1993]. The algorithm simply counts the co-occurrence frequencies between source language words and target language words. This process yields the measure $P(W_{source} \wedge W_{target})$ which is then used, in conjunction with $P(W_{source})$ and $P(W_{target})$ to estimate the conditional probability $P(W_{target}|W_{source})$ (what [Brown et al.1993] call the *language model probability.* The conditional probability measure is then used to construct a dictionary for the source and target languages.[1]

[1] Because our goal is to produce a bilingual lexicon, we do not generate individual term alignments per sentence, although such a procedure can be done during term alignment processing.

The approach developed in this paper uses an estimate of word frequency in this way, but adds information about the syntactic category of each word(*a tag*) in a sentence. It will be shown that by using this type of linguistic information to filter the space of possible correlations for a given term, the process of learning word alignments is more accurate and requires less training.

The Alignment Problem

Sentence Alignment

The problem of aligning sentences within a document has been the focus of two different classes of algorithms. In [Gale and Church 1991] and [Brown et al.1991] a dynamic programming method for probabilistic alignment is introduced. In these systems a document-wide sentence mapping is created, correlating source and target language sentences by their sentence lengths. They rely on the assumption that sentences of similar length and relatively close in position are likely to translations of each other.

Methods based on sentence length heuristics have performed well when clean documents are properly segmented at *hard boundaries.*[2] In fact, error rates of 4.62% when aligning segmented documents have been reported [Gale and Church 1991]. However, for very large corpora, in which data may not be reliably segmented or reliably translated (i.e. sentence/paragraph omissions, mis-translations, etc.), such systems report that error rates increase threefold [Gale and Church 1991].

Newer methods do not rely on the existence of *hard boundaries* to perform accurately. Systems such as [Melamed 1996] and [Simard et al.1992] rely on generating *bitext maps*, character level mappings between source and target language documents. These methods use similarity measures at the level of words [Melamed 1996] and character ([Simard et al.1992] and [Church 1993]) to generate points of correspondence in a *bitext space*. [Melamed 1996] and [Simard et al.1992] then calculate best-fit linear approximations to find a bitext map. These approaches have proven to be highly robust as they are able to process imperfect documents without sacrificing accuracy.

TAG-ALIGN uses an extended sentence alignment method similar to [Gale and Church 1991], while incorporating features of the word/character similarity found in [Melamed 1996], [Simard et al.1992] and [Church 1993]. Instead of relying on raw sentence length alone, TAG-ALIGN incorporates word alignments generated by its term alignment module in its measure of sentence similarity.

Word Alignment

Given a sentence in a source language and its translation in a target language, a word alignment is a mapping between each a word of a source sentence and zero or more corresponding words in the target sentence. Mappings to target words need not be unique. Thus multiple source words may map to a single target word. A typical mapping is shown in figure 1:

Figure 1 illustrates the mapping between a sentence in English and its translation in Irish Gaelic. It should be noted that the mappings are very often non-linear and unordered as languages may differ in the word order, whether they use determiners (i.e. *the, a, an, these*), and various other ways. These differences make it difficult to construct algorithms that accurately build word level alignments.

[2] *In [Gale and Church 1991] these are paragraph boundaries.*

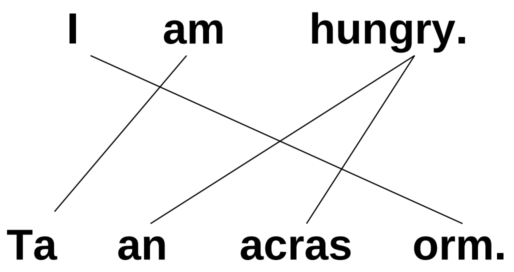

Figure 1: *A Word Alignment for Irish Gaelic and English*

Current systems fall into two categories: alignment based on word coocurrence in sentences, and alignment based upon character level similarities between words in documents. Algorithms of the former category attempt to estimate conditional probabilities for a target word's occurrence given the presence of a source word within a section of a document (see [Brown et al.1993]). These systems can be considered adaptive, as they attempt to learn these conditional probabilities from training data.

Approaches of the latter category rely on similarities in the orthography of words (cognates) to build *anchor* points in a document. These anchors are then used to compute a best-fit linear approximation for an overall document alignment at the character level (see [Melamed 1996] and [Simard et al.1992]) which can then be used to create a word level alignment. Because these methods generally do not require documents to be sectioned, they are quite robust in their ability to handle poorly sectioned data. However, because they attempt to find linear approximations, they are less likely to create accurate sentence alignments like those shown in figure 1.

Methods that rely on word coocurrence attempt to learn mapping rules between source language and target language words. As such they are more apt to handle word order variations between dissimilar languages. These systems act, essentially, as Bayesian classifiers, attempting to discover word(s) in a target language that belong to a class designated by a source language word. These classes can be seen as bilingual dictionaries giving source language words and their possible translations in a target language.

However, coocurrence alone is inadequate for obtaining reliable dictionaries of this sort. The problem lies in the fact that within sentences of a given language, two words may collocate frequently. Consider the sentence, "The police arrested him at midnight." In this case one finds that the word *police* is highly likely to occur with the *arrest*. Thus a coocurrence based classifier might learn erroneous mappings of this sort: $policia \rightarrow police, arrest$. Because word occurrences in human languages are often *not* independent distributions, Bayesian methods are likely to hypothesize too many such mappings of this type.

The TAG-ALIGN System

The TAG-ALIGN system is designed to provide sentence and word level alignments. Figure 2 provides a dataflow diagram of TAG-ALIGN's different functional modules. The sections below describe the operations of each of these individual modules.

The overall alignment process begins with the Sentence Alignment module, which generates aligned sentence pairs from bilingual input texts.

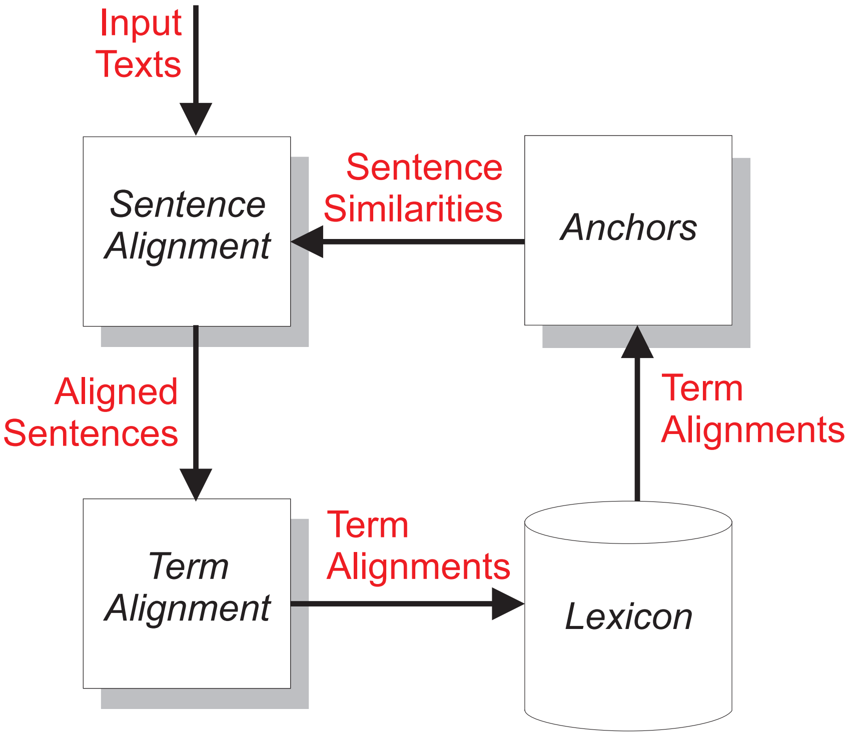

Figure 2: *Block Diagram of tag-align*

These sentence pairs are passed to the Term Alignment module which generates lexical entries for each source word encountered in its input and stores each entry into its lexicon. These entries are subsequently used by the Anchors module to evaluate the similarity of new sentences that are to be aligned by the sentence aligner.

Tag-based Word Alignment

The TAG-ALIGN approach presented here borrows from the translation models of [Brown et al.1991] by using conditional probabilities to find likely translations of a source word in target sentences. However, it improves upon these approaches through its use of linguistic filters (discussed below).

Initially, TAG-ALIGN acts much like a Bayesian classifier. Source/target sentence pairs are used to tabulate word correspondences. For example, if a word w_s appears in the source sentence, and a word w_t appears in the target sentence, the coocurrence frequency $w_s \wedge w_t$ is incremented. Given this frequency and the frequencies of w_s and w_t we can calculate an estimate of the coocurrence probability as given by the (1).

$$P(w_s \wedge w_t) = \frac{|w_s \wedge w_t|}{|w_t| + |w_s|} \tag{1}$$

where $|w_t|$ denotes the frequency of w_t

For the purposes of building a bilingual lexicon, it is useful to calculate the conditional probability given in (2).

$$P(w_s|w_t) = \frac{|w_s \wedge w_t|}{|w_s|} \tag{2}$$

By maximizing this conditional probability over possible target words in a sentence, it is possible to determine a likely mapping between source words and target words (in effect, creating a bilingual dictionary between the source and target languages).

The equations given in (1) and (2) provide a simple Bayesian selection process for independent w_t. Unfortunately, words that are members of a sentence are related to each other by syntactic rules making the assumption of independence questionable. This leads simple Bayesian classification to generate spurious correlations.

To see this problem, one need only consider the correspondence between nouns and determiners. It is rare in English (and many other languages) for determiners to appear in a sentence without a noun. As such, a source sentence noun $noun_s$ is likely to cooccur with two words in the target sentence: $article_t$ and $noun_t$.

To alleviate this problem, the TAG-ALIGN algorithm uses syntactic part of speech tags (e.g. noun, adjective, verb, etc) for words in the source and target sentences to eliminate improbable correspondences between source and target words. Thus a verb in the source sentence $verb_s$ is not considered to correspond to an article $article_t$ in the target language. In essence, part of speech (POS) tags are used to filter the space of possible target correlates for a source word.

TAG-ALIGN processes sentence pairs by tabulating a word coocurrence probabilities for every possible source/target word pair allowed by the POS mapping rules described in table 1.

Rules for which parts of speech to consider as valid correspondences vary for different source/target language pairs. Table1 shows mapping rules for part of speech tags between Spanish(source) and English(target).

These rules are motivated by the intuition that within language pairs, the possibilities of certain POS tag mappings are highly improbable while others are more probable. For the Spanish/English rules given above, one finds that wh-elements in the source language (Spanish) can only correlate with other wh-elements in the target language. In essence, these POS mapping rules act to constrain the hypothesis space of target language words.

Ideally, these rules would be represented as a probability distribution describing likelihoods for all source/target mapping between POS tags. This way a more accurate description of POS mapping rules could be expressed. Thus the conditional measure of coocurrence for a source/target term pair could be given by equation 3.

$$Align(w_s, w_t) = P(w_s|w_t)P(Tag_s|Tag_t) \tag{3}$$

Discrete distributions of $P(Tag_s|Tag_t)$ could be empirically measured by examining aligned bilingual corpora.

The rules given in Table 1 are an ad hoc estimate of these discrete distributions. Once, large sets of word-aligned sentences can be obtained, it will be possible to integrate more accurate empirical distributions into TAG-ALIGN.

By using this added information, it is possible to eliminate spurious classifications like those described above. Part of speech tagging is done as a preprocessing step to source and target language sentences of the input document set.

Table 1: *POS Mapping rules from Spanish to English*

Spanish	V	Adj	Adv	N	Pn	Pnt	P	Exp	D	Conj	Neg	Pron	Num	WH
English	V Adv	Adj Adv	Adv Adj	N	Pn	Pnt	P	Exp	D	Conj	Neg	Pron Pn	Num	WH

Table 2: *Part of Speech Tags used in TAG-ALIGN*

Tag Names	**Part of Speech**
V	Verbs
Adj	Adjectives
Adv	Adverbs
N	Nouns
Pn	Proper Names
Pnt	Punctuation
P	Prepositions
Exp	Expletives (e.g. There)
D	Determiners
Conj	Conjunctives (e.g. and, or, that, which)
Neg	Negation (e.g. no and not)
Pron	Pronouns
Num	Numbers
WH	WH-element (e.g. who, when, where, etc)

Feedback-based Sentence Alignment

As stated in the introduction, TAG-ALIGN is capable of using alignments learned during term alignment to generate better sentence alignments. In practice, TAG-ALIGN uses its generated lexicon to measure similarity between sentences, an *anchor* score. An anchor score for a sentence pair is calculated by equation 4. The anchors function yields a maximum of 1.0 when all translations of a every source word are found in the target sentence, since the value $\sum_{w_s} P(w_s|w_t) = 1.0$.

$$anchors(s_s, s_t) = \frac{\sum_{w_t \in s_t} \sum_{w_s \in s_s} P(w_s, w_t)}{|s_s|} \tag{4}$$

Essentially, equation 4 provides a quantification of the number of words that are similar between two sentences. When a correlation cannot be found for a source/target word pair (i.e. the value $P(w_t|w_s)$ is unavailable) a cognate score is computed using LCSM (Longest Common Substring Match), which weights words by the similarity of the character sequence of which they are composed.

The anchor function is used along with sentence length are used as similarity functions for a sentence alignment protocol similar to [Gale and Church 1991] and [Brown et al.1991] with certain enhancements. TAG-ALIGN performs a dynamic programming search for the best alignment allowing source/target sentence pairs to be matched, inserted, deleted, joined, and split. For further discussion of these processes, see [Gale and Church 1991].

Discussion

Term Alignment

We ran an initial set of experiments to evaluate the TAG-ALIGN algorithm. These experiments compared TAG-ALIGN against a baseline Bayesian classifier that simply counts word coocurrence and source/target word frequencies. The resulting bilingual lexicons were then compared for accuracy.

Both systems were presented with text data from the proceedings of the United Nations during the 1992 year. These data exist in parallel texts, translated by human translators. They comprise more than 650 documents.

Words within a document are then stemmed to common root forms. This process joins two inflected forms on the same word to a common root (for example, *running, ran* and *run* all become *run*). Stemming helps to prevent probability mass from being distributed across variants on a single word.

Once stemming has been done, sentences are tagged with POS information. Then a process of aligning sentences is undertaken. Of the 654 documents aligned, over seventeen megabytes of data were extracted. Sentence pairs with POS tags are then presented to the two word alignment algorithms for processing.

Both the baseline classifier and TAG-ALIGN's Term Alignment module were able to learn a number of interesting term translations. However, TAG-ALIGN was able to postulate many more correlations that the Bayesian system could not. Some of these are given in Table 3.

Spanish	**English**	$P(S\|E)$
Academia	Academy	0.857
absorber	absorb	0.800
multa	fine	0.670
saludar	address	0.275
optar	choose	0.333

Table 3: *Postulated term mappings with their probability scores*

Interestingly, the baseline classifier was unable to find most mappings. It seems that without tags, a Bayesian classification system would require much more data than our rather small 17 megabyte corpus was able to provide. TAG-ALIGN was able to postulate more correlates for the same corpus as evidenced by Table 3.

A more graphic illustration of the effect of tagging can be seen in Table 4. Here the Spanish word *poste* (post/station) is deemed by the baseline system to correlate with a series of proper nouns. The TAG-ALIGN system is able to eliminate all of these as irrelevant leaving only English nouns as possible correlates. In this case, spurious mappings are eliminated by POS tags.

There were certain instances where tags are unable to provide the necessary distinction. An example of this is shown in the raw data in Figure 3. Here, the verb *oprimir* (to oppress) is

Method	Spanish Term	English Term	$P(S\|E)$
Baseline	poste	Chai Gharehtappeh	1.000
		Changuleh	1.000
		Dush K	1.000
		Fakkeh	1.000
		Iraqi	0.080
		Khosravi	1.000
		Khosrawi	0.500
TAG-ALIGN	poste	border	0.171
		bunker	0.600
		coordinate	0.127

Table 4: *Mapping for the Spanish word "poste"*

```
oprimir 0.00001269 +open (liberate 0.16667) (oppress 0.66667)
```

Figure 3: *Raw data generated by TAG-ALIGN for the Spanish word "oprimir"*

postulated to map to *liberate* (with a score of 0.166667) and *oppress* (with a score of 0.66667). Because *oppress* and *liberate* frequently cooccur in sentences like *"It is right to liberate those who are oppressed."*, and because both words are verbs, TAG-ALIGN is unable to rule out *liberate* as a possible translation of *oprimir* without more data.

Sentence Alignment

The lexicon generated by TAG-ALIGN using the 1992 UN proceedings was used to align UN data from 1993. This corpus, composed of 42 files. None of these files were contained paragraph boundary information making it difficult for a [Gale and Church 1991] type aligner to recover from localized alignment mistakes (i.e. omissions, mistranslations, etc.).

TAG-ALIGN's Anchor-based sentence alignment module was run with and without the acquired lexicon the results were compared for accuracy. Table 5 shows the performance gain when the acquired lexicon is used.

Algorithm	**Accuracy**
TAG-ALIGN w/ lexicon	88.1%
TAG-ALIGN w/o lexicon	73.8%

Table 5: *Sentence Alignment Performance w/ and w/o an acquired lexicon*

Additionally, code from [Gale and Church 1991] was run (results not reported here). But the accuracy was severely limited because certain transformations are not modeled by [Gale and Church 1991].[3]

Conclusion

Tag-based alignment can help reduce the hypothesis space of a simple Bayesian classification. By using linguistic filtering, it is possible to achieve high accuracy rates with less training data. As more data becomes available, it may be possible to refine these results and achieve even higher levels of accuracy. Furthermore, data acquired during tag-based alignment can be used to generate more accurate sentence alignments.

Thus it is conceivable that a more sophisticated system could be developed to learn better rules for POS tag mappings. By closely examining the relationships between source and target languages, it might be possible derive probabilistic mapping rules that would give more a accurate discriminator for the source/target word correlation process.

A promising avenue of inquiry may be to apply these concepts to other existing alignment schemes similar to [Melamed 1996] and [Simard et al.1992]. [Melamed 1996] reports that his SIMR/GSA system already benefit from having a lexicon. If a feedback mechanism could be implemented with such a system, even more robust and accurate sentence/word alignments might result.

References

[Fung and Church1994] P. Fung and K. W. Church, 1994 K-Vec: A New Approach for Aligning Parallel Texts *Proceedings of the 15th International Conference on Computational Linguistics*, Kyoto, Japan.

[Gale and Church1991] W. Gale and K. W. Church 1991 A Program for Aligning Sentences in Bilingual Corpora *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, Berkeley, CA.

[Brown et al.1993] P. Brown, S. Della Pietra, V. Della Pietra, and R. Mercer 1993 The Mathematics of Machine Translation: Parameter Estimation *Computational Linguistics*, 19:263–312

[Church1993] K. W. Church 1993 Char_align: A Program for Aligning Parallel Texts at the Character Level *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, Columbus, OH.

[Brown et al.1991] Peter F. Brown, Jenifer C.Lai, and Robert Mercer. 1990. Aligning Sentences in Parallel Corpora *Computational Linguistics*, 16:79–80

[Melamed1996] I. Dan Melamed 1996 A Geometric Approach to Mapping Bitext Correspondence *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Philadelphia.

[3] *Their algorithm considers insertion, deletion, 1-1, 2-2, (1-2), (2-1), and crossover mappings, but does not consider N to 1 or 1 to N mappings for arbitrary N, although TAG-ALIGN does.*

We believe this performance degradation is primarily due to the lack of paragraph boundaries and the high number of N to 1 mappings in the UN data. This data irregularity does not seem to manifest as frequently in other UN data sets.

[Simard et al.1992] Michel Simard, G. F. Foster and P. Isabelle, 1992 Using Cognates to Align Sentences in Bilingual Texts *Proceedings of TMI-92*, Montreal, PQ.

[Dagan et al.1993] I. Dagan, K. Church, and W. Gale 1993 Robust Word Alignment for Machine Aided Translation *Proceedings of VLC-93*, Columbus, Ohio.