From Word Alignment to Machine Translation via Superlinks

Lars Ahrenberg
Department of Computer and Information Science
Linköpings universitet
S-58183 Linköping
lah@ida.liu.se

Håkan Jonsson
Nordisk Språkteknologie A/S
P.O. Box 93
N-5701 Voss
Hakan.Jonsson@gri.no

Abstract

This paper presents a data-driven lexicalist framework for machine translation based on alignment. In addition to word alignment that provides the word correspondences of source and target language, the system uses classifications of the correspondences based on supertags. We first give an overview of the framework, pointing out its fundamental concepts and then report some results from a pilot implementation and evaluation on the ATIS domain.

1. Introduction

Automatic word alignment systems have reached a level of performance that make them useful for a number of applications including machine translation.

A problem with many current systems, though, especially when applied in the construction of machine translation systems, is that they align surface strings with no information on properties such as part-of-speech or sense. Moreover, they do not provide information on the context of a word correspondence. Unless this information is provided somehow, it is impossible to select the correct alternative translation(s) for a given context. To illustrate, consider the following extract from a (constructed) parallel text:

(a) They will win. *De kommer att vinna.*
(b) Did they win? *Vann de?*
(c) They did win. *De vann faktiskt.*
(d) They never win. *De vinner aldrig.*

Given more data of this kind a word alignment system would be able to find the lexical correspondences *win:vinna, win:vann*, and *win:vinner*. Moreover, it would also find correspondences such as *they:de* and *will:kommer_att*. These correspondences do not suffice, however, to determine the proper translation of *win* for any of the English sentences above, as there is no information at hand to guide the choice.

Methods for statistical machine translation such as Brown et al. (1990, 1993) have some measures at their disposal to handle alternatives in translation. First, the dictionary is probabilistic so that every alternative translation is assigned a probability. But consistent selection of one alternative will obviously give the wrong alternative in many cases, so these probabilities would have to be conditioned by contextual factors. One possibility is a bigram language model for the target language. Another possibility employed by these early approaches were to make translation contingent on alignment of word positions. However, none of these methods is sufficient to differentiate the translations in pairs (b) and (c) above.

In this paper we present a way of representing and processing word correspondences with their local context that enables their immediate use in a system for automatic translation. We refer to this representation as a *superlink* (cf. Ahrenberg, 2000) and the method of translation as *Superlink Constrained Lexical Transfer (SCLT)*. The paper is organised as follows: Sections 2 and 3 provide an overview of the framework. Section 4 presents an initial experiment of constructing a translation system from a translation corpus of the ATIS-domain (Jonsson, 2000).

2. Supertags and superlinks

A superlink is, basically, a pair of supertags, where one element of the pair refers to the source language and the other element refers to the target language. A supertag is "a rich description of a lexical item that impose complex constraints in a local context" (Bangalore and Joshi, 1999 : 237).

While Bangalore and Joshi use the LTAG formalism in their work, the formalism of the supertags as well as their scope can be chosen in different ways. The tags used in our work so far have a more limited power of expression.

We make a general distinction between *inherent* tags and *relational* tags. A relational tag includes an argument, while an inherent tag does not. The inherent tags used here bear strong resemblance to traditional POS-tags, but have in some of our experiments been augmented with semantic properties.

TABLE 1. **Some inherent tags from the English ATIS tagset**

Tag	Meaning	Examples
DIS	Det, indefinite, singular	"a"
NAS	Common noun, singular	"flight"
PMS	Proper noun, city	"Boston"
PSS	Pronoun, subject, singular	"I", "he"
SS	Preposition, spatial	"from"
VIB	Verb, indefinite base form	"need"

TABLE 2. **Possible relational tags in the SCOTS notation**

Tag	Meaning
LeftOf(X)	item is to the left of argument X.
LeftAdj(X)	item is left-adjacent to argument X.
RightOf(X)	item is to the right of argument X.
RightAdj(X)	item is right-adjacent to argument X.
First	item is first in the sentence.
Last	item is last in the sentence.

A few examples of inherent tags are given in Table 1, while the full set of relational tags used in the evaluation are given in Table 2. Note that relational tags are restricted to a few linear relations between words in a string. The argument X of the relational tags can be of three kinds: a token, an inherent tag or a disjunctive set of inherent tags. A supertag in our framework, then, is a conjunctive set of inherent and relational tags. An example of a superlink fitting the word correspondence *win:vinna* in sentence (a) is <VIB&RightAdj(VR) :: VIB&RightOf(VFR)>.

3. A translation model based on superlinks

Now, given the notion of a superlink we can give a characterization of the translation relation between the sentences of two languages for a given text type:

A sentence T is a translation of a (well-formed) sentence S iff

- There is a set of token alignments $\langle s_i, t_i \rangle$ such that $S' = s_1\ s_2\ \ldots\ s_n$ is a proper tokenization of S and T is a well-formed permutation of $T' = t_1\ t_2\ \ldots\ t_n$, where some of the t_i may be null tokens.
- For every pair $\langle s_i, t_i \rangle$ there exists a superlink $\langle ST_i, TT_i \rangle$ such that (i) ST_i is a proper supertag for s_i, and (ii) TT_i is a proper supertag for t_i.

This definition of the translation relation decomposes it into three relations: Thus, S is related to S' via tokenization, S' to T' via transfer, and T' to T via transposition (or permutation). However, as the transfer phase is based on the superlinks we must introduce the supertags in the processing somehow; this is done in a separate monolingual tagging phase. Finally, the transposition phase was extended with a phase of output formating to simplify the treatment of compounds, altogether giving a translation process in five stages: tokenization, tagging, transfer, transposition and output formatting.

The superlinks employed in the transfer stage act as filters on the set of translations for a given word. Probabilities may also be used at this stage. The transposition stage reorders target language words so as to satisfy constraints associated with target language supertags. Finally, output formatting performs tasks such as capitalization and compounding.

The framework can be described as lexicalist and data-driven. It is lexicalist in the sense that all processing operates on strings of lexical items (tokens). It is data-driven since a large part of the knowledge is supposed to be derived from translation corpora.

The most important knowledge sources in the system are the following:

- A source dictionary SDIC associating source tokens with supertags for the source language,
- A target dictionary TDIC associating target tokens with supertags for the target language,
- A transfer dictionary WLINKS associating source tokens with target tokens,

- A superlink dictionary SLINKS associating souce language supertags with target language supertags.

Every dictionary in this list may be obtained automatically from a translation corpus, or by generalizing results obtained from such a corpus. The source and target dictionaries are obtained from tagged versions of the translation corpus. The transfer dictionary WLINKS is obtained from (possibly corrected) links generated by a word alignment program. The superlink dictionary can be obtained in the same way as the transfer dictionary by looking at the tags rather than the words.

The central idea for transfer in this model is that the set of potential translations of a source token is filtered by the superlinks that are compatible with the actual local context. Hence the name Superlink-Constrained Lexical Transfer (SCLT).

The result of lexical transfer is a bag of pairs <token, supertag> for the target language. When several source tokens have alternative translations that are jointly allowed by WLINKS, TDIC and SLINKS, the number of bags can become quite large. In the prototype system all bags were in fact multiplied out and the set was pruned when a certain threshold had been reached.

The rules concerned with target language word order can take on a variety of forms. The general idea is to find those sequences that satisfy the constraints given by the target language supertags.

4. Evaluation

The purpose of the evaluation was to create a prototype system that could serve as a proof-of-concept for the proposed method. This prototype we refer to as SCOTS, Superlink COnstraint Transfer lexicalist translation System.

Some specific questions were:

- What could be achieved by creating all necessary translation data automatically? In particular, how would this system compare with a baseline system doing word-for-word translation based on most frequent alignments for tokens of the English source file?
- What would be the difference in performance between automatic versions and versions where translation data are improved manually?
- Would the system, in any version, be able to reproduce the corpus, or produce translations of the same standard as those in the corpus?

The overall strategy was to run and test SCOTS in a variety of different conditions. Many potential variables could be manipulated to create different testing conditions but full and systematic variation of these variables is not possible. A selection of test cases had to be made, where some of these variables were explored. In order to do this, three test series were devised.

In the first test series tag sets were varied on both the English and the Swedish side. Here though, we only report results for the most elaborated tag set. Automatic alignment was also contrasted with manual alignment. All supertags were trigram tags, i.e. every word is given information about its inherent tag and the inherent tags of its two nearest neighbours. For instance, the word *Boston* occuring in the context *'from Boston to'* would be assigned the trigram tag (PMS, RightAdj(SS), LeftAdj(SS)).

In the second test series the knowledge sources were improved manually after inspection of system errors. The third test series was performed to investigate some issues that had arised during the first two series. For lack of space, we cannot report all of the evaluation here, but give only a brief summary.

The corpus chosen for the experiment was a small English-Swedish corpus of machine-translated sentences from the ATIS domain, where the translations were produced by the SLT system (Agnäs et al., 1995). This corpus was chosen partly because we wanted to see whether the SCLT method could produce results similar to a knowledge-based MT system. A few sample translations from the corpus are given in Figure 1.

The corpus was divided into a training set consisting of 232 sentence pairs and a test set comprising 30 pairs.

In all test cases translation was performed in the direction from English to Swedish.

does continental fly from denver to san francisco
flyger continental airlines från denver till san francisco

what ground transportation is available in boston.
vad finns det för marktransport i boston.

i would like a flight from philadelphia to dallas.
jag skulle vilja ha en flygning från philadelphia till dallas.

Figure 1. Example translations in the ATIS corpus.

4.1 Extraction and construction of translation data

The process of extracting translation information was divided into three stages. The stages are ordered in such a way that the actions in each stage are prerequisite to carry out the next.

In stage 1, the preprocessing stage, tokens are created. A phrase extraction tool (Merkel & Andersson, 2000) was used to retrieve recurrent multi-word units (MWUs) automatically from both the English and the Swedish halves of the ATIS corpus.

Then the tokenized English and Swedish input files were tagged using a Brill tagger with the different tag sets. From this output each token was assigned a trigram tag.

In stage 2, required files were automatically generated using different combinations of tokens and tagged language files. All such generations were only done using the training corpus. Several different versions of each of the four required domain files, TDIC, SDIC, SLINKS and WLINKS were produced. The SDIC and TDIC files are created from the tagged texts. The automatic WLINKS file is created with the use of of a word alignment system (Ahrenberg et al, 2000). This program does not only generate a lexicon but also link instances which are used in combination with the tagged files to create different SLINKS files. A sample entry from the SDIC files is the following:

```
<boston,
        ((PMS, RightAdj(SS), Last)[0.35],
        (PMS, RightAdj(SS) LeftAdj(S))[0.04],
        (PMS, RightAdj(SS), LeftAdj(SS))[0.54],
        (PMS, RightAdj(SS), LeftAdj(ST))[0.08])>
```

The figures associated with the trigram tags reflect are probabilities estimated by Maximum Likelihood Estimation from the corpus. They are primarily used for rating and filtering the word bags that result from transfer.

In stage 3, the best automatically extracted files were improved manually, e.g. by restructuring compounds, generalizing tags across natural word classes, introducing relational tags that refer to tokens rather than tags, and more.

4.2 Evaluation criteria and measures

The following criteria were chosen as a basis for evaluation of performance:

- *Word Selection Criteria*. Ability to select the correct TL tokens. Measured as recall and precision.
- *Word Order*. Ability to order the target language tokens correctly. Measured for correct target tokens only.
- *Identity*. Identical reproduction of the reference translations
- *Translation* quality, subjectively experienced. Two measures were used: accuracy and grammaticality.

4.3 Results and discussion

A sample of the measures obtained are given below. In Table 3 the different test cases are explained, while Table 4 provides the figures. The baseline case (Bl) means word-for-word translation based on most frequent translation for any given word.

TABLE 3. **Test Cases Shown**

Test case	SL inh. tagset	TL inh. tagset.	Word Al-ignment	Super-links
1:3	ATIS	ATIS	Auto	Auto
1:6	ATIS	ATIS	Manual	Auto
2:4	ATIS	ATIS	Manual	Refined
3:1	As 2:4 + lexical coverage of test corpus			
Bl	-	-	Manual	-

TABLE 4. **A sample of results (Tr means training corpus, Tst test corpus)**

Case	Pcn	Rcl	Ord	% Id	% Gr	Acc
1:3 Tr	.809	.855	.740	30.6	45.0	.700
1:3 Tst	.687	.739	.679	0.0	20.0	.463
1:6 Tr	.935	.950	.953	66.8	95.0	.950
1:6 Tst	.802	.825	.789	16.7	65.0	.788
2:4 Tr	.995	.992	.998	95.3	90.0	1.00
2:4 Tst	.791	.811	.796	16.7	60.0	.663
3:1 Tst	.855	.855	.822	33.3	50.0	.700
Bl Tr	.813	.846	.748	13.8	-	-
Bl Tst	.791	.817	.761	10.0	-	-

These measures warrant the following conclusions:

- There is clear evidence (1.6 vs. 1.3) that performance improves as word alignment gets more accurate. The quality of the word alignment is, not surprisingly, very important for the quality of the output.
- Word-by-word translations achieve rather high evaluation scores. However, translation using

automatically extracted superlinks performs better than manually refined word-by-word translation. There is a marked difference for the training corpus, but only a slight difference for the test corpus.

- With manual refinements, the ceiling of all measures are approached on the training corpus. It is quite possible on this domain to make SCOTS memorize the whole training corpus.
- On the other hand, results on the test corpus are markedly lower than on the training corpus. These results must be seen against the background of the sparse training corpus causing the test corpus not to be well covered by the superlinks generated from the training corpus.
- The performance on the training corpus increases from case 1:6 to case 2:4. Further results shows that this increase was largely due to the addition of proper handling of compounds. This increase in performance is not visible on Accuracy, however.
- The scores of case 3:1 are a marked improvement to those of 2:4 on the test corpus. Still, the scores are lower than what is achieved on the training corpus.

As the corpus used has been small and the ATIS domain is simple and restricted, we are not in a position to draw definite conclusions as to the performance of the method. However, we regard these results as encouraging.

There are obviously a number of ways in which to extend and improve this work. First of all, the SCOTS prototype should be tested on a larger and more complex corpus. Second, it would be interesting to study the effects of using more informative supertags, e.g. syntactic information in the form of dependency relations between words.

References

Agnäs, M-S., Alshawi, H., Bretan, I., Carter, D., Ceder, K., Collins, M., Crouch, R., Digalakis, V., Ekholm, B., Gambäck, B., Kaja, J., Karlgren, J., Lyberg, B., Price, P., Pulman, S., Rayner, M., Samuelsson, C., & Svensson, T. (1994). *Spoken Language Translator: First-Year Report. 1994*. Swedish Institute of Computer Science, SICS. SICS research report R94:03. ISRN: SICS-R-94/03-SE.

Ahrenberg, L, Andersson, M. and M. Merkel (2000). A knowledge-lite approach to word alignment. In J. Veronis (ed.) Parallel Text Processing : Alignment and Use of Translation Corpora, pp. 97-116. Kluwer Academic Press.

Bangalore, S. And A. K. Joshi, 1999. Supertagging: An Approach to Almost Parsing. *Computational Linguistics* 25(2): 237-265.

Brown, Peter, John Cocke, Stephen Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, John D. Lafferty, Robert L. Mercer and Paul S. Roossin, 1990. A Statistical Approach to Machine Translation. *Computational Linguistics*, 16(2): 79-85.

Brown, Peter F., S. A. Della Pietra, V. J. Della Pietra and R. L. Mercer, 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2): 263-311.

Jonsson, Håkan (2000). Exploring Superlink Constrained Lexicalist Transfer in Machine Translation. Master's thesis, Department of Computer and Information Science, Linköpings universitet.

Merkel, M., & Andersson, M. (2000). Knowledge-lite extraction of multi-word units with language filters and entropy thresholds. *Proceedings of the 6th Conference on Content-Based Multimedia Information Access, RIAO-2000*, Paris, France.