A Corpus-Based Approach to Deriving Lexical Mappings

Mark Stevenson
Department of Computer Science,
University of Sheffield,
Regent Court, 211 Portobello Street,
Sheffield S1 4DP
United Kingdom
marks@dcs.shef.ac.uk

Abstract

This paper proposes a novel, corpus-based, method for producing mappings between lexical resources. Results from a preliminary experiment using part of speech tags suggests this is a promising area for future research.

1 Introduction

Dictionaries are now commonly used resources in NLP systems. However, different lexical resources are not uniform; they contain different types of information and do not assign words the same number of senses. One way in which this problem might be tackled is by producing mappings between the senses of different resources, the "dictionary mapping problem". However, this is a non-trivial problem, as examination of existing lexical resources demonstrates. Lexicographers have been divided between "lumpers", or those who prefer a few general senses, and "splitters" who create a larger number of more specific senses so there is no guarantee that a word will have the same number of senses in different resources.

Previous attempts to create lexical mappings have concentrated on aligning the senses in pairs of lexical resources and based the mapping decision on information in the entries. For example, Knight and Luk (1994) merged WordNet and LDOCE using information in the hierarchies and textual definitions of each resource.

Thus far we have mentioned only mappings between dictionary senses. However, it is possible to create mappings between any pair of linguistic annotation tag-sets; for example, part of speech tags. We dub the more general class *lexical mappings*, mappings between two sets of lexical annotations. One example which we shall consider further is that of mappings between part of speech tags sets.

This paper shall propose a method for producing lexical mappings based on corpus evidence. It is based on the existence of large-scale lexical annotation tools such as part of speech taggers and sense taggers, several of which have now been developed, for example (Brill, 1994)(Stevenson and Wilks, 1999). The availability of such taggers bring the possibility of automatically annotating large bodies of text. Our proposal is, briefly, to use a pair of taggers with each assigning annotations from the lexical tag-sets we are interested in mapping. These taggers can then be applied to, the same, large body of text and a mapping derived from the distributions of the pair of tag-sets in the corpus.

2 Case Study

In order to test this approach we attempted to map together two part of speech tag-sets. We chose this form of linguistic annotation because it is commonly used in NLP systems and reliable taggers are readily available.

The tags sets we shall examine are the set used in the Penn Tree Bank (PTB) (Marcus et al., 1993) and the C5 tag-set used by the CLAWS part-of-speech tagger (Garside, 1996). The PTB set consists of 48 annotations while the C5 uses a larger set of 73 tags.

A portion of the British National Corpus (BNC), consisting of nearly 9 million words, was used to derive a mapping. One advantage of using the BNC is that it has already been tagged with C5 tags. The first stage was to re-tag our corpus using the Brill tagger (Brill, 1994). This produces a bi-tagged corpus in which each token has two annotations. For example `ponders/VBZ/VVZ`, which represents the token is `ponders` assigned the Brill tag `VBZ` and `VVZ` C5 tag.

The bi-tagged corpus was used to derive a pair of mappings; the word mapping and the tag mapping. To construct the word mapping from the PTB to C5 we look at each token-PTB tag pair

and found the C5 tag which occurs with it most frequently. The tag mapping does not consider tokens so, for example, the PTB to C5 tag mapping looks at each PTB tag in turn to find the C5 tag with which it occurs most frequently in the corpus. The C5 to PTB mappings were derived by reversing this process.

In order to test our method we took a text tagged with one of the two tag-sets used in our experiments and translate that tagging to the other. We then compare the newly annotated text against some with "gold standard" tagging. It is trivial to obtain text annotated with C5 tags using the BNC. Our evaluation of the C5 to PTB mapping shall operate by tagging a text using the Brill tagger, using the derived mapping to translate the annotations to C5 tags and compare the annotations produced with those in the BNC text. However, it is more difficult to obtain gold standard text for evaluating the mapping in the reverse direction since we do not have access to a part of speech tagger which assigns C5 tags. That is, we cannot annotate a text with C5 tags, use our mapping to translate these to PTB tags and compare against the manual annotations from the corpus. Instead of tagging the unannotated text we use the existing C5 tags and translate those to PTB tags. Each approach to producing gold standard data has problems and advantages. The Brill tagger has a reported error rate of 3% and so cannot be expected to produce perfectly annotated text. However, when we tag the text with PTB tags and use the mapping to translate these taggings to C5 annotations we have no way to determine whether erroneous C5 tags were produced by errors in the Brill tagging or the mapping.

Our test corpus was a text from the BNC consisting of 40,397 tokens. Both word and tag mappings were created in each direction (PTB to C5 and C5 to PTB). To apply the tag mapping we simply used it to convert the assigned annotation from one tag-set to the other. However, when the word mapping is applied there is the danger that a word-tag pair may not appear in the mapping and, if this is the case, the tag mapping is used as a default map.

The results from our evaluation are shown in Table 1. We can see that the C5 to PTB word mapping produces impressive results which are close to the theoretical upper bound of 97% for the task. In addition the word mapping in the opposite direction is correct for 95% of tokens.

Although the results for the word mappings in each direction are quite similar, there is a significant difference in the performances of the default mappings, 86% and 74%. Analysis suggests that the PTB to C5 default mapping is less successful than the one which operates in the opposite direction because it attempts to reproduce the tags in a fine-grained set from a more general one.

Type	Direction	
	C5 to PTB	PTB to C5
Word	97%	95%
Tag	86%	74%

Table 1: Mapping results

3 Conclusion and Future Work

This paper considered the possibility of producing mappings between dictionary senses using automatically annotated corpora. A case-study using part of speech tags suggested this may be a promising area for future research.

Our next step in this research shall be to extend our approach to map together dictionary senses. The reported experiment using part of speech tags assumed a one-to-one mapping between tag sets and, while this may be reasonable in this situation, it may not hold when dictionary senses are being mapped. Future research is planned into ways of deriving mappings without this restriction. In addition, we will also explore methods for deriving mappings when corpus data is sparse.

References

E. Brill. 1994. Some advances in transformation-based part of speech tagging. In *AAAI-94*, Seattle, WA.

R. Garside. 1996. The robust tagging of unrestricted text: the BNC experince. In J. Thomas and M. Short, editors, *Using corpora for language research: Studies in Honour of Geoffrey Leach.*

K. Knight and S. Luk. 1994. Building a large knowledge base for machine translation. In *AAAI-94*, Seattle, WA.

M. Marcus, B. Santorini, and M. Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Tree Bank. *Computational Linguistics*, 19.

M. Stevenson and Y. Wilks. 1999. Combining weak knowledge sources for sense disambiguation. In *IJCAI-99*, Stockholm, Sweden. (to appear).