THE EXPLOITATION OF CORPORA IN NATURAL LANGUAGE PROCESSING

Anvarova Sarvinoz Jumanazar qizi,
sarvinozanvarova97@gmail.com
Mirzo Ulug‘bek nomidagi O‘zbekiston milliy universiteti magistranti

Annotation: One of the first things required for natural language processing (NLP) tasks is a corpus. In linguistics and NLP, **corpus** (literally Latin for body) refers to a collection of texts. Such collections may be formed of a single language of texts, or can span multiple languages -- there are numerous reasons for which multilingual corpora (the plural of corpus) may be useful. Corpora may also consist of themed texts (historical, Biblical, etc.). Corpora are generally solely used for statistical linguistic analysis and hypothesis testing.

Key words: *NLP, part-of-speech tagging, CLAWS, disambiguation, annotation, human intuition, language analysis, evaluation, testbed corpora.*

NLP is a rapidly developing area of study, which is producing working solutions to specified natural language processing problems. The application of annotated corpora within NLP to date has resulted in advances in language processing-part-of-speech taggers, such as CLAWS, are an early example of how annotated corpora enabled the development of better language processing systems (see Garside, Leech, and Sampson 1987). Annotated corpora have allowed such developments to occur as they are unparalleled sources of quantitative data. To return to CLAWS, because the tagged Brown corpus was available, accurate transition probabilities could be extracted for use in the development of CLAWS. The benefits of this data are apparent when we compare the accuracy rate of CLAWS-around 97 per cent to that of TAGGIT, used to develop the Brown corpus-around 77 per cent. This massive improvement can be attributed to the existence of

annotated corpus data which enabled CLAWS to disambiguate between multiple potential part-of-speech tag assignments in context.

It is not simply part-of-speech tagging where quantitative data are of prime importance to disambiguation. Disambiguation is a key problem in a variety of areas such as anaphor resolution, parsing, and machine translation. It is beyond doubt that annotated corpora will have an important role to play in the development of NLP systems in the future, as can be seen from the burgeoning corpus-based NLP literature (LREC 2000).

Beyond the use of quantitative data derived from annotations as the basis of disambiguation in NLP systems, annotated corpora may also provide the raw fuel for various terminology extraction programs. Work has been developed in the area of automated terminology extraction which relies upon annotated corpora for its results (Daille 1995; Gausier 1995). So although disambiguation is an area where annotated corpora are having a key impact, there is ample scope for believing that they may be used in a wider variety of applications.

A further example of such an application may be called evidence-based learning. Until recently, language analysis programs almost exclusively relied on human intuition in the construction of their knowledge/rule base. Annotated corpora corrected/ produced by humans, while still encoding human intuitions, situate those intuitions within a context where the computer can recover intuitions from use, and where humans can moderate their intuitions by application to real examples. Rather than having to rely on decontextualized intuitions, the computer can recover intuitions from practice. The difference between human experts producing opinions about what they do out of context and practice in context has long been understood in artificial intelligence-humans tend to be better at showing what they know rather than explaining what they know, so to speak. The construction of an annotated

corpus, therefore, allows us to overcome this known problem in communicating expert knowledge to machines, while simultaneously providing testbeds against which intuitions about language may be tested. Where machine learning algorithms are the basis for an NLP application, it is fair to say that corpus data are essential. Without them machine learning-based approaches to NLP simply will not work.

Another role which is emerging for the annotated corpus is as an evaluation testbed for NLP programs. Evaluation of language processing systems can be problematic, where people are training systems with different analytical schemes and texts, and have different target analyses which the system is to be judged by. Using one annotated corpus as an agreed testbed for evaluation can greatly ease such problems, as it specifies the text type/types, analytical scheme, and results which the performance of a program is to be judged upon. This approach to the evaluation of systems has been adopted in the past, as reported by Black, Garside, and Leech (1993), for instance, and in the Message Understanding Conferences in the United States (Aone and Bennett 1994). The benefits of the approach are so evident, however, that the establishment of such testbed corpora is bound to become increasingly common in the very near future.

One final activity which annotated corpora allow is worthy of some coverage here. It is true that, at the moment, the range of annotations available is wider than the range of annotations which it is possible for a computer to introduce with a high degree of accuracy. Yet by the use of the annotations present in a hand-annotated corpus, a resource is developed that permits a computer, over the scope of the annotated corpus only, to act as if it could perform the analysis in question. In short, if we have a manually produced treebank, a computer can read the treebank and discover where the marked constituents are, rather than having to work it out for itself. The advantages of this are limited yet clear. Such a use of an annotated corpus may provide an economic means of evaluating whether the development of a certain

NLP application is worthwhile-if somebody posits that the application of a parser of newspaper stories would be of use in some application, then by the use of a treebank of newspaper stories they can experiment the worth of their claim without actually producing a parser.

There are further uses of annotated corpora in NLP beyond those covered here. The range of uses covered, however, is more than sufficient to illustrate that annotated corpora, even though we can justify them on philosophical grounds, can more than be justified on practical grounds.

REFERENCES

1. Aone, C. and S. W. Bennett. 1994. `Discourse tagging and discourse tagged multilingual corpora: Proceedings of the International Workshop on Sharable Natural Language Resources (Nara), 71-7.

2. Daille, B. 1995• Combined Approach for Terminology Extraction: Lexical Statistics and Linguistic Filtering. Unit for Computer Research on the English Language Technical Papers 5, Lancaster University.

3. Garside, R., G. Leech, and A. M. McEnery. 1997. Corpus Annotation. London: Longman.

4. G. Sampson. 1987. The Computational Analysis of English. London: Longman.

5. Leech, G. 1997. `Introducing corpus annotation: In Garside, Leech, and McEnery (1997),1-18.

6. LREC 2000. Proceedings of the 2nd International Conference on Language Resources and Evaluation (Athens).