11-712: NLP Lab Report

Jesse Dodge

April 25, 2014

Abstract

[one paragraph here summarizing what the paper is about –NAS]

[brief introduction –NAS]

1 Basic Information about French

French is a Roman language, spoken in many countries around the world. It's the 10th most commonly spoken language in the world, with between 220 and 300 million people speaking French as a first or second language. French syntax is quite similar to English, with a few key differences. For example, most French adjectives go after the word they're modifying, the French negation has two parts, and determiners have gender.

2 Past Work on the Syntax of French

French syntax has been well studied. There exist texts ranging from simple (e.g. Fundamentals of French Syntax[1]) to complex (e.g. The Syntax of French[2], Foundations of French Syntax[3]). Beyond just textbooks, there are a number of papers that address individual topics in French such as verb movement[4] and tokenization[5].

With repsect to syntax for parsing, there exists a French constituency-parsed treebank[6] which contains 12,531 sentenecs from Le Monde, a French newspaper. This treebank has been annotated in a similar style to the Penn treebank, with a few additions. For example, each word (when applicable) contains information on gender.

Note: we can put something more here about other syntax formalisms as they come up.

3 Available Resources

TreeTagger[7] is an open-source POS tagger for French. Europarl[8] has 194 MB copus of French-English text that I can annotate. Simply taking the first 2000 words and dividing it into A and B test sets will give us the data we need. Note: The fact this is parallel data is interesting, but not

[1] http://www.academia.edu/1997083/Fundamentals_of_French_Syntax
[2] http://www.amazon.com/Syntax-French-Cambridge-Guides/dp/B008SLJ1WQ
[3] http://www.amazon.com/Foundations-French-Cambridge-Textbooks-Linguistics/dp/0521388058
[4] http://minimalism.linguistics.arizona.edu/AMSA/PDF/AMSA-175-1000.pdf
[5] https://giguete.users.greyc.fr/pricai96/part4.html
[6] http://alpage.inria.fr/statgram/frdep/fr$_s tat_d ep_p arsing.html$
[7] http://www.cis.uni-muenchen.de/ schmid/tools/TreeTagger/
[8] http://www.statmt.org/europarl/

necessary. (Note: There exists a large corpus of Le Monde[9] that I can use either as the test corpora as well, if Europarl is too formal. Also, this data can be used for semi-supervised training, though it is annonated already with constituency parses.)

4 Survey of Phenomena in [Your Language/Genre –NAS]

5 Initial Design

6 System Analysis on Corpus A

7 Lessons Learned and Revised Design

8 System Analysis on Corpus B

9 Final Revisions

10 Future Work

References

[9] http://alpage.inria.fr/statgram/frdep/fr$_{s}tat_{d}ep_{p}arsing.html$