Annotating Signs of Syntactic Complexity to Support Sentence Simplification

Richard Evans (R.J.Evans@wlv.ac.uk) and Constantine Ouseen (C.Ouseen@wlv.ac.uk)

Syntactic complexity

- Key aspects of syntactic complexity:
  - Coordination
    - She knew the risks and still insisted the operation should go ahead, Dr Addicott said.
  - Subordination
    - McKay, of Wark, Northumberland, denies five charges of contaminating food.

- Signs of syntactic complexity:
  - Coordinators (conjunctions and punctuation marks)
  - Subordination boundaries (complementisers, wh-words, and punctuation marks)

- Annotation scheme encodes:
  - linking functions of coordinators
  - bounding functions of subordination boundaries
  - information about false signs that do not have a coordinating or bounding function

Research context: the FIRST project

- Aims to improve accessibility of text for readers with autistic spectrum disorders (ASD). This process includes syntactic simplification.
- Syntactic simplification can be framed as the conversion of long syntactically complex sentences into sequences of shorter sentences that convey the same meaning.
- It can improve:
  - the reliability of NLP applications (IE, MT, and syntactic parsing)
  - the accessibility of written texts for people with reading difficulties.
- We are developing a system for syntactic simplification which operates in two steps:
  1. Identifying and classifying signs of syntactic complexity
  2. Rewriting sentences using manually crafted rules

Signs of syntactic complexity

1. Coordinators (linking functions)

- Example: CLN
  - "We had a stable [and] loving family"

2. Subordination boundaries (bounding functions)

- Example: CMA1
  - "But the melancholy experience of is quite a thing as well! Thee dear knows what I mean" Theodor W. Adorno

Advantages of the scheme

- Fine grained annotation enables flexibility in the design of syntactic simplification algorithms
- Penn Treebank (Marcus et al., 1993):
  - Does not directly encode specific linking/bounding functions of conjunctions and commas.
  - Expensive to produce.
- Van Delden and Gomez (2002)
  - Focus only on classification of commas.
  - Expensive to produce.
- Maler et al. (2012)
  - Requires existing syntactic annotation.

Corpus analysis 1

Sign and class distribution

- Most frequent disagreements
  - News (K=0.70)
  - Healthcare (K=0.74)
  - Literature (K=0.76)

- All annotations: K=0.77

Corpus analysis 2

Inter-annotator agreement

- Annotations using this scheme:
  - Reproducible
  - Annotation of the full syntactic structure of each sentence is not necessary.
  - Consistent
  - Non-experts can annotate texts from specialist domains in a way that is reliable and consistent
  - Expenditure
  - The resources have been successfully used to train an NL classification of signs of syntactic complexity (Dornescu et al., 2013)

Future work:

- Identification of additional markables (hyphens, parenthetical and class labels SSMI)
- Exploitation of the resources in NLP applications

Acknowledgements

The research described in this paper was partially funded by the European Commission under the Seventh Framework Programme for Research and Technological Development (FP7, ICT 2011.5.1 FIRST 287607). This publication reflects the views only of the Authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.