SHORT TERM SCIENTIFIC MISSION (STSM) – SCIENTIFIC REPORT

The STSM applicant submits this report for approval to the STSM coordinator

Action number: IS1312
STSM title: Automatic Disambiguation of Multi-sense Discourse Connectives
STSM start and end date: 18/02/2018 to 24/03/2018
Grantee name: Maja Buljan

PURPOSE OF THE STSM/

The aim of the STSM was to approach the task of disambiguating multi-sense discourse connectives from a linguistic and statistical perspective. Focusing on the contextual ambiguities of discourse adverbials in English, the effect of usage context on discourse relationships was to be modelled by features. The purpose of these features was to create a model for automatic labelling of discourse connectives' arguments, depending on the sense ambiguity and complexity of the examined connectives.

An attempt to automate discourse relation labelling was deemed beneficial both for providing insight into the capabilities and drawbacks of statistics-based language models, and direction for future focused linguistic analysis. Additionally, an STSM focused on analysing discourse relations through several methods would provide the necessary background and experience to initiate such an investigation for other languages, where working groups had not yet undertaken a project of such scale or depth regarding the same question.

DESCRIPTION OF WORK CARRIED OUT DURING THE STSM

After examining the supervisors' previous work on the senses and usage of discourse connectives, the STSM began with an in-depth investigation of available corpora, and patterns of behaviour of the discourse connectives in focus. By analysing small to medium-sized manually annotated corpora, simple rules were developed to enable the use of large, superficially annotated corpora for the training and testing of an argument-labelling model.

The subject of focus was the PLF (Practical Lexical Function) model of composition, based on the principles of distributional semantics. While the PLF model has so far predominantly been used for modelling content words, one of the research questions was: how well would the PLF model perform when modelling function words, specifically discourse connectives?

Four discourse connectives were chosen, varying in the breadth of senses and inter- vs. intra-sentential distribution of arguments: *while*, *since*, *otherwise*, and *instead*. The working problem was formulated as: given the discourse connective, its sense, and its second argument, is it possible to predict the connective's first argument in unannotated data?

A simple rule-based approach was used to extract training data from a large-scale corpus – a concatenation of MASC, BNC, ukWaC, and Wikipedia – while data from the annotated PDTB was used as the gold-standard test set. Models

were built for the four selected discourse connectives, with the additional training of two separate models for *since* – the causal and the temporal sense of the connective.

The research question was tested by composing the discourse connective and its second argument, and then performing another composition with the candidates for the first argument. When defining the scope of the composition, several levels of complexity were evaluated – the argument clause head, as a lemma vector; the head as a function; the entire clause as an additive composition; the entire clause as a functional composition.

The viable first arguments were selected by observing the candidates' projection in the vector space, and their similarity or distance from the projection of the gold set – representing a cluster to which the first arguments should be expected to be assigned to.

DESCRIPTION OF THE MAIN RESULTS OBTAINED

The experiments demonstrated the capabilities and drawbacks of automated connective and argument labelling. With regards to the applicability of a CDSM (compositional distributional semantics model), the PLF model proved to be suited to the task of modelling discourse connectives. The success of the model (measured in vector similarity between the projected arguments and the gold set arguments) depends, predictably, on the breadth of context given when modelling the argument clauses, with similarity increasing as the clause vector moved from individual head lemmas to composed phrases, reflecting the semantics of the clause as a unit.

A clear difference emerged between the difficulty of modelling particular connectives from the chosen four. Modelling *otherwise* was not as straightforward as the other connectives, given its three distinct senses which could not be easily differentiated in the unannotated data. The average difference between the positive and negative clustering results (vector similarity to cluster centroid) for otherwise was 34%. *Instead* and *since* proved much more discriminative, the former averaging a 57% difference, and the latter averaging 56% for the temporal sense, and 84% for the causal.

The most difficult connective to model proved to be *while*, with an average difference of 19% between the positive and negative clustering results. This is most prominently due to the widely varying meanings of the surface lemma *while*, which introduced noise in the training data. Given that the training data only had an automatic POS-level annotation (which was also observed to have errors in manual inspection of the corpus), acquiring correct examples of *while* as a discourse connective, along with its first and second clausal argument, proved difficult with only a handful of parsing rules.

In conclusion, the experiment showed that applying CDSMs to discourse relations is promising, but that quality annotated data is, as always, paramount to building a robust and reliable model.

FUTURE COLLABORATIONS (if applicable)

As the initial results proved promising, there is an interest on both sides to delve deeper into the compatibility of CDSMs and discourse. Supervised work on this topic will continue, with the intention of exploring the possibilities in more detail, with a larger collection of annotated corpora, and more focused experiments.