SHORT TERM SCIENTIFIC MISSION (STSM) – SCIENTIFIC REPORT

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

Action number: *IS1312*
STSM title: "The path to hearer-old status: Modeling how entities enter common sense knowledge"
STSM start and end date: 15/01/2018 to 06/04/2018
Grantee name: Ieva Staliūnaitė

PURPOSE OF THE STSM/

This STSM had two main objectives:

- Learning more about hearer-old and hearer-new information status distinctions in referrring expressions
- Building a tool that could aid automatic summarization of documents that are written some time prior to the summarization

These two goals included the subgoals of identifying people and organizations in a large news corpus, linking these entities across documents, recognizing the type of modifiers that the referring expressions contained, defining linguistic and distributional features of the referring expressions, using the metadata of the artciles to distinguish different kinds of entities, building a classifier to predict the information status of an entity in the future, analyzing the features that change as entities become hearer-old.

DESCRIPTION OF WORK CARRIED OUT DURING THE STSMS

During this STSM I used the New York Times annotated corpus for analyzing how descriptions of entities change with time. I worked on extracting noun phrases that refer to named entities (together with the types of modifiers used with them) and linking them within and across documents with the help of automatic parsing and coreference resolution as well as corpus metadata. Furthermore, I collected the topic and desk tags as well as the timestamps at the article level. I used word embeddings to measure similarity between the different topics and clustered the similar topics into broader topic groups.

With that information and the parses of the descriptions themselves, I defined features that could be expected to have an effect on the information status of that entity at the time. Some examples are:

- how many verbs do the descriptors contain

- how many possessives do the descriptors contain

- how much of a variation in words in the descriptors is there (type/token ratio)

- how many definite articles are there

- how many of the mentions have appositives

- how many honorifics are used

- how many of the mentions have 1-3 word descriptors

- how many of the mentions have 20 or more word descriptors

- how many mentions fall in articles about topic X

- how many mentions have relative clause descriptors

- how many mentions have pre-modifiers

- a score evaluating how big the gaps between the mentions within the month are

I then tested whether these features are relevant to the prediction of the time that it will take for the entity to become well-known. Furthermore, I built a model that predcits whether an entity will be accepted as common knowledge or not after a period of time.

DESCRIPTION OF THE MAIN RESULTS OBTAINED

The research carried out in this STSM provides insight into the features that are relevant when deciding on the information status of an entity. Some of the features define the kind of an entity that it is (such as topic, frequency and honorifics used). The significance of such features in the prediction shows that different entities have different patterns of evolving into well known objects. In addition to it, features such as the use of posessives, determiners and appositives show the linguistic patterns that people adopt when talking about hearer-new or hearer-old entities.

The most important results of the analysis of variance show that descriptions of hearer-old organizations use more definite articles than hearer-new organizations, which is consistent with the theory of informaiton status, which claims that information status affects the definiteness of a referential expression. In addition, definite phrases referring to people chage from the use of definite article to the use of posessives, which is consistent with psycholinguistic research which shows that people tend to omit the less informative words (such as articles) when the sentence they are producing is less dense with information.

Moreover, the results indicate that appositives are used to describe people who are important but have only recently started appearing in the news, and thus are the ones that can be predicted to become well known in the near future.

FUTURE COLLABORATIONS (if applicable)

We are planning on continuing the work on this topic, as many questions were raised while trying to resolve the initial hypotheses. To begin with, thinking of different ways to define acceptance has been part of this project and we believe that other deifnitions may produce better results. For instance, we could define an entity not as universally accepted as hearer-old, but only as accepted with regard to a particular aspect. For instance, while everyone might know Facebook as the social media company, perhaps not everyone thinks of it as an advertising mogul. Thus in a context about ads the author of the article might want to inform the readers of this aspect of the entity.

In addition, we want to expand this research beyond only people and organizations, to nouns that refer to everyday objecst and concepts that are formed new and accepted into the common knowledge over a period of time. For instance, it would be interesting to see how modern technology is incorporated into the common ground, such as the emergence of the Internet, cryptocurrency, selfies and other new things that become very common.

Furthermore, we would like to continue working on a sequence to sequence model, which would produce the referring expressions as they would look like after a certain period of time has passed. We would use what we learnt about the important factors in order to produce the training data.