**RESEARCH PROPOSAL**

**Michalina Skibicka**

## 1. Short description of the project

Research proposal submitted by the candidate Michalina Skibicka concerns the study of human annotation in Machine Learning/Computational Linguistics tasks based on the task of similarity scoring in sets of paraphrases.

The study would aim to observe if the score in inter-annotator agreement is dependable on whether the annotators follow the guidelines provided or not and whether introducing automatic pre-annotation based on the chosen linguistic tools such as Universal Dependencies parsing, Part-Of-Speech tagging and scoring-related labels could lead to annotation quality improvement and to partial elimination of wide scoring differences between annotators in the case of categorically identical sentences.

The study would also aim at providing additional information on the nature of a given degree of similarity in a paraphrase and on how to construct ideally correct paraphrases of varying degree of similarity, i.e. the aim would be to look for patterns of how a given degree of similarity in a paraphrase could be achieved, which could potentially be used for the construction of an automatic paraphrase generator that would enable the improvement of automatic creation of adequately diverse train/test datasets for various tasks in Machine Learning and Natural Language Processing.

## 2. Position of the project within the discipline

The Machine Learning boom that is being currently experienced both in the commercial and in the academic world creates a need for new Artificial Intelligence tools and thus for the endless iterations of the training and testing processes that require an increasingly large amount of carefully prepared, adequately diverse and adequately repetitive data.

As to the current candidate's commercial experience, it seems like most of the projects make use of human labour to expand ML datasets and corpora: this means creating a myriad of similar tasks for human linguists who are required to produce a number of paraphrases to a given sentence or simply to a product feature each time a new feature is introduced. Candidate's work as an annotator on Samsung's team winning SEMEVAL 2016 contest solution (description of the solution by its authors available here: https://www.aclweb.org/anthology/S/S16/S16-1091.pdf) and the further cooperation with one of its authors (K. Chodorowska, currently at Interdisciplinary Centre for Mathematical and Computational Modelling, University of Warsaw) led her to two main observations:

1. Certain linguistic in nature and repeatable patterns of features which were manipulated by the linguists to create paraphrase sets of required diversity can be discerned.
2. The differences in similarity scoring of categorically identical sets of sentences seem to follow certain patterns as well: there exist given linguistic and discursive text properties that make human annotators score similar sets of sentences differently, regardless of the instruction provided, which, in the case of training sets, introduces an inconsistency error for the machine and hinders similarity recognizers' performance.

In the case of the first observation, high repeatedness of such features seems to provide relatively good grounds for trials of automatization of paraphrase generation. The second observation invites a study that could lead to findings that would be helpful in increasing human annotator agreement by partially eliminating the factor of linguists not following the instructions provided by task makers in similarity scoring tasks. Such findings seem to have substantial potential of fitting into the niche of

automatic paraphrase generation and to contributing to the solution to the problem of human annotation quality.

## 3. Short description of the data

Initial research on both parts of the project has been done based on the publicly available SEMEVAL 2016 corpora (http://alt.qcri.org/semeval2016/task1/index.php?id=data-and-tools) and on machine translation of the Europarl Parallel Corpus (http://www.statmt.org/europarl/). Since the corpora are domain-specific and differ in construction, it seems that each corpora presents different difficulties.

The candidate has prepared the preliminary analysis of the characteristics of the corpora and has discerned a number of factors that contribute to the overall scoring difficulty of these corpora, such as:

- informational density of the text,
- quantity of modifiers,
- spoken syntax in written text,
- level of noun and modifier abstractiveness,
- frequency of rhetoric means used compared to the length of a sentence,
- sentence length in tokens, etc.

The next step would be the analysis of the paraphrase set aimed at discerning which properties could be manipulated to construct a research paraphrase corpora to be used in the scoring experiment, and, potentially, in case if they would prove not to have a significant effect on scoring – for considerations on automatic paraphrase generation. Initial observations led the candidate to choose a number of linguistic properties that were related to the scoring instructions provided in the course of the SEMEVAL task and that could be manipulated to obtain model paraphrases of a given degree (on a 0 to 5 scale in SEMEVAL 2016 task), which could include:

- domain-specific synonymy,
- grammatical tense change,
- voice change,
- proportion of missing/additional modifiers,
- grammatical number change,
- sentence length compared to the compositionality of sentence meaning,
- metaphor use, etc.

The research proposed by the candidate could be based on the same corpora and make use of the preliminary findings presented above, and the very experiment could be replicated on a resulting corpora that would be prepared by the candidate herself on the basis of the SEMEVAL corpora.

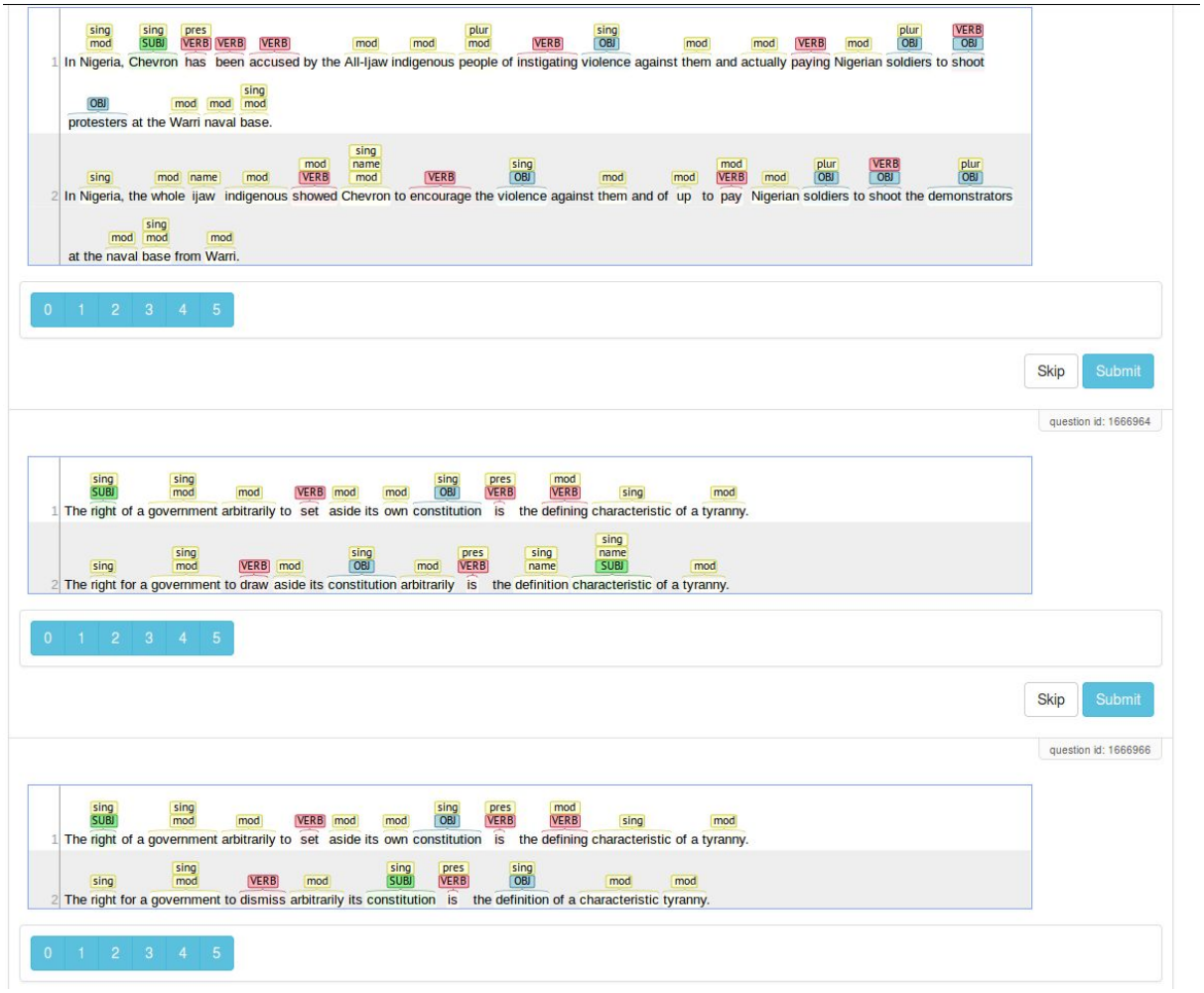
The number of entries in a research corpora could be determined in the course of the project, yet it is estimated that a minimum number of sentences in a set would be ca. 1000 original sentences and 5 paraphrases to an original sentence in order for such a set to be usable for Machine Learning purposes.

## 4. Short description of the project's research method

The project could be built on the basis of the 5-step similarity scoring provided by the SEMEVAL competition or another graded similarity scoring scale. The workflow presented below uses the SEMEVAL scale but can be fully adjusted to another method. The workflow of the project as proposed is as follows:

1. Analysis of the existing corpora (for example SEMEVAL corpora) focusing on finding corpora properties that may introduce possible scoring difficulties for human annotators. The analysis would consist of extracting the examples of such properties, setting them together and calculating their frequency in the existing corpora.

2. Defining a set of paraphrase set properties that would be tested in the course of the experiment based on the analysis of the existing corpora and their current scorings: such characteristics should be prepared for each score separately based on the scoring instruction.
3. Creating a new corpora for the purpose of the experiment that would mirror the analysis of the existing corpora. Research corpora could also be obtained through modification of the existing corpora in order to cover the properties to be tested. It is assumed that the repartition of scores should be balanced, i.e. in a corpus of 1000 entries with a set of 5 paraphrases to each original sentence, each paraphrase would receive respectively the scores 0, 1, 2, 3, 4, 5.
4. Tools to be used in the experiment would be: POS tagger, morphosyntactic dictionary and UD parser, with the possible addition of SemGrex and TokensRegex from Stanford Core NLP Demo package. The desired output of these tools would be mapped to the tested properties in order to create labels for pre-annotation. The annotation tool of choice would be the BRAT annotation tool (http://brat.nlplab.org/) since it allows various modifications in labelled pre-annotations. Example of the prototype of the prepared task is presented below:




The labels above the sentences are related to some of the tested properties: for example a number of modifiers the change in which is attributed to various score changes according to the SEMEVAL task guidelines is displayed. The label *mod* that covers the modifier was mapped to the output of

additionally trained Stanford Core NLP POS tagger, namely to the tags of: adjective, modifier and adverb.
The sentences here are displayed in pairs: original above, paraphrase below, although they can also be displayed in a full set with original sentence and its five paraphrases beneath it.
The 0-5 button below the sentences is a scoring button to be used by the linguists to score paraphrases in this task.
The platform used to produce the example is a the LangTasks.com platform: http://freelancer.langtask.com/?ckattempt=1

5. A test round of scoring without annotation would be conducted, scores then counted and evaluated according to their correctness and consistency with the instructions to linguists.
6. A round of scoring with pre-annotation would be conducted, its scores counted and evaluated according to the same criteria as the set without annotation.
7. Scores in both rounds should be compared, comparison for each tested property should be recorded. Conclusions of the comparison should be used to confirm or refute research hypothesis if human annotation quality and inter-annotator agreement could be improved by the use of NLP tools based pre-annotation.

## 5. Short statement on the relevance of the project

Initial academic research shows high relevancy of the project, as such hypothesis has not been tested before (to candidate's best knowledge). The research shows significant potential to innovate: it might provide knowledge to be used in building products needed on the market such as automatic paraphrase generators, it fits the research niche and may lead to the improvement of the current annotation standards.

## 6. Statement of the provisional timetable

1. Preparation of corpora and the task: **1 – 2 months** (depending on the platform chosen to prepare the task)
2. Annotation: **1 month**

With the assumption that an experienced annotator has the average scoring speed of 800 sentence pairs per day (8 hours) and the inexperienced annotator of 200 to 300 sentence pairs per day, and given the complexity and novelty of the task, it seems to be reasonable to take the 200 sentence scores per day as expected. In the case of a corpus composed of 1000 entries with 5 paraphrases the number of pairs to be scored is 5000, which would mean that an inexperienced annotator would need 25 working days to complete the task. Such working time seems a safe assumption given the candidate's previous experience.

3. Analysis of the results: **1-2 months**
4. Additional work: **to be determined**