1. Introduction

Automatic text simplification is the process of reducing the complexity of a text to make it more accessible to a broader range of readers with different readability levels. While preserving its meaning as much as possible, a text’s lexical, syntactic and discourse level features should be made more simple. However, evaluating the simplicity level of a text is still a challenging task for both humans and automatic systems.

Current approaches to automate text simplification vary depending on the type of simplification. Lexical simplification was the first effort in this area. In particular, Devlin and Tait (1998) introduced an approach of replacing words with their most common synonym based on frequency (Kučera et al., 1967). More recently, publicly available resources such as Simple English Wikipedia and the Google 1T corpus have been used to automate lexical simplification based on similar approaches such as common synonym replacement and context vectors (e.g. Biran et al., 2011; Bott et al., 2012; Rello et al., 2013; Kauchak, 2013).

Another approach to automatic text simplification involves syntactic simplification. Current work in this area aims to identify and simplify complex syntactic constructions such as passive phrases, embedded clauses, long sentences, etc. Initial work on syntactic simplification focused on the use of transformation rules in order to generate simpler sentences (e.g. Chandrasekar and Srinivas, 1997). Later, work have paid more attention on sentence splitting (e.g. Carroll et al., 1998), rearranging clauses (e.g. Siddharthan, 2003) and dropping clauses (e.g. (Barlacchi and Tonelli, 2013; Štajner et al., 2013)). To our knowledge, Siddharthan (2003) is the only effort that specifically addressed the preservation of a text’s discourse structure by resolving anaphora and ordering sentence.

In the remainder of this paper, we describe the methodology we used to measure the 4 simplification criteria of the QATS workshop: GRAMMATICALITY, MEANING PRESERVATION, SIMPLICITY and OVERALL. In Sections 2 and 3, the details of our submitted system are described, while Section 4 summarises our results.

2. System Overview

As can be seen in Figure 1, our system consisted of three independent supervised models in order to predict each of the three main aspects: GRAMMATICALITY, MEANING PRESERVATION and SIMPLICITY. We used 10 fold cross-validation in order to choose the best supervised models. The 4th aspect (i.e. OVERALL) was predicted using the predictions of MEANING PRESERVATION and SIMPLICITY.

2.1. Grammaticality Prediction

In order to predict the quality of the simplified sentences from the point of view of grammaticality, we have used the log likelihood score of the sentences using the Google Ngram corpus. To do this, the BerkeleyLM language modeling toolkit was used (Pauls and Klein, 2011) to built a language model from the Google Ngram corpus, then the perplexity of all simple sentences in the training set were calculated. These log likelihood scores were used as features to feed a Random Forest classifier.

2.2. Meaning Preservation Prediction

The purpose of meaning preservation is to evaluate how close the meaning of the original sentence is with respect to its simple counterpart. To do this, we used two complementary approaches based on word embedding and the cosine measure.

2.2.1. Word Embedding

We used the Word2Vec package (Mikolov et al., 2013a) to learn the representation of words on the Wikipedia dump. We then trained a skip-gram model using the deeplearning4j library. As a result, each...
word in the original sentence and its simple counterpart are represented as a vector. As calculating the similarity of two sentences using word embedding is still a challenging task, our approach to this problem was to use average similarity. To do so, we calculated the similarity of each word (each vector) in the original sentence to all the words in its simpler counterpart. Then using the average length of the original and simple pairs, we calculated the average similarity between a pair of sentences. This similarity was the first feature we fed to a Random Forest classifier.

2.2.2. Cosine Similarity of WordNet Synonyms

The second feature we used for meaning preservation was based on WordNet synonyms. Each sentence was represented as a vector of its constituent words. Then, using WordNet all synonyms of each word were added to the corresponding vector of the sentence. However, as each word can have various part of speech (POS) tags, before expanding the vector, we first identified the POS of all the words in the sentence using the Stanford POS tagger (Manning et al., 2014). Afterwards, we filtered the synonyms according to the POS tags and added only those with the same POS tag of the word. As a result, each sentence was represented as a vector of words and their synonyms. Using the cosine similarity to calculate the similarity between corresponding vectors of pairs of sentences, we measured how close the meaning of two sentences were.

\[
\text{Cosine Sim}(i) = \frac{O_i \cdot S_i}{||O_i|| \times ||S_i||} \tag{1}
\]

2.3. Simplicity Prediction

The purpose of simplicity prediction is to evaluate how simpler the simple sentences are compared to their original counterpart. As simplicity can be measured at various levels (i.e. lexical, syntactic and discourse), we considered the following sets of features in order to capture the changes at each level.

2.3.1. Vector Space Model Similarity

The first feature we considered in order to evaluate the simplicity of the simple sentences compared to their original counterpart, was the cosine similarity between the Term Frequency-Inverse Document Frequency (TF-IDF) vectors of each pair. A cosine similarity of 1 indicates that no change has been made in the simplification process. However, before transforming sentences into their corresponding TF-IDF vector, we preprocessed them. First, stop words were removed, then all words were stemmed using the Porter Stemmer (Porter, 1980). As a result, each sentence was represented as a vector of the size of all the stems in all sentences. It is worth noting that in order to compute the inverse document frequency for each stem, we considered each sentence as a document. The cosine similarity between original and simple sentences of the \(i^{th}\) pair is calculated using Formula 1, where \(O_i\) and \(S_i\) represent the vectors of the original sentence and its simple counterpart correspondingly.

2.3.2. Sentence Length

Traditional approaches to readability level assessment have identified text length as an important feature to measure
complexity (e.g. Kincaid et al. (1975)). Following this, we investigated the influence of sentence length in terms of the number of open class words only. By ignoring closed class words, we eliminated the effect of words which do not contribute much to the meaning of the sentence. Thus, we considered the difference between the length of pairs of sentences as our second feature for simplicity prediction.

2.3.3. Average Word Length

According to Kincaid et al. (1975), not only can the number of words in the sentence be an indicator of simplicity level, but also its length in terms of the number of characters. To account for this, we also considered the difference between the average number of characters between pairs of sentences. Using this feature along with the number of words of each sentence (see Section 2.3.2), we investigated not only the influence of the length of sentence, but also the length of each word in the sentence.

2.3.4. Frequency in the English Wikipedia Corpus

The frequency of each word in the regular English Wikipedia can be an indicator of the simplicity level of the word. We expected that words in the original sentences would be more frequent in the regular English Wikipedia than words of the simple sentences. Thus, we calculated the difference between the average frequency of all words of the original sentence and their simple counterpart. To do this, we preprocessed both pairs of sentences and the regular English Wikipedia corpus, in order to remove stop words and then stem the remaining words.

2.3.5. Frequency in the Simple English Wikipedia Corpus

The Simple English Wikipedia corpus[^1] is an aligned corpus of 60K <regular, simple> pairs of Wikipedia articles. We used this corpus in order to calculate the average frequency of words of each pair of sentences. We expected the words of simpler sentences to be more frequent in the Simple English Wikipedia articles compared to the original sentences. To do this, we performed the same preprocessing as described in Section 2.3.4 and used the average frequency of the sentence’s stems as features.

2.3.6. Frequency of Cue Phrases

The last feature we considered to predict the simplicity aspect was the difference in the usage of cue phrases. Cue phrases are special terms such as however, because, since, etc. which connect text segments and mark their discourse purpose. Several inventories of cue phrases have been proposed (e.g. Knott, 1996; Prasad et al., 2007)). For our work, we used the list of 100 cue phrases introduced by Prasad et al. (2007) and calculated the difference between the frequency of cue phrases across pairs of sentences. It is worth noting that cue phrases may be used to explicitly signal discourse relations between text segments or may be used in a non-discourse context. However, here we considered both discourse and non-discourse usage of cue phrases.

[^1]: http://www.cs.pomona.edu/~dkauchak/simplification/
| System Name                          | Accuracy | MAE  | RMSE |
|-------------------------------------|----------|------|------|
| Grammaticality-Davoodi-RF-perplexity| 58.73%   | 27.38| 34.66|
| Meaning preservation-Davoodi-RF     | 49.21%   | 30.56| 36.31|
| Simplicity-Davoodi-RF-0.5           | 34.92%   | 43.25| 45.52|
| Simplicity-Davoodi-RF-0.6           | 35.71%   | 41.67| 44.48|
| Simplicity-Davoodi-RF-0.7           | 35.71%   | 40.48| 44.48|
| Overall-Davoodi-0.5                 | 34.15%   | 41.27| 44.21|
| Overall-Davoodi-0.6                 | 32.54%   | 42.06| 45.67|
| Overall-Davoodi-0.7                 | 33.33%   | 41.27| 45.16|

Table 2: Official Results of our system at QATS.