1. Introduction

NLP can be used in several ways to help people read texts that are inaccessible to them in some way. Machine translation can be used to translate texts from a language one does not know, and summarization can be used to shorten texts in order to make them faster and/or easier to read. Both these techniques do in a sense improve readability. In this paper, we describe a small initial study of the interplay between readability, summarization, and MTranslatability. We applied an extractive summarizer to a set of news stories, and investigated the effect on readability and machine translation (MT).

Extractive summarizations are created by extracting the most important sentences from a text; no re-writing of sentences takes place. Smith and Jönsson (2011a) showed that by using extractive summarization, readability was improved according to several readability measures. Their study was performed on Swedish text, and showed an effect on several genres, including news.

MTranslatability is the translatability of a text by an MT system, i.e., how easy a text is for an MT system to translate (Bernth and Gdaniec, 2001). This notion has mostly been exploited in connection to rule-based MT systems, for instance by improving MT output through re-writing the source according to some controlled language rule set (see e.g., Roturier (2004)). MTranslatability is also related to the notion of confidence estimation (CE), to estimate the quality of machine-translated sentences (Specia et al., 2009). Many features used for CE are also similar to those used in many readability formulas, such as sentence length. CE differs from the notion of MTranslatability, however, in that in CE the estimation is done after the translation phase, whereas MTranslatability is related to assessing whether a sentence or text is easy or hard to translate before it is sent to the MT system.

2. Experiment

We performed an experiment where we applied an automatic summarizer to English news text, which was then translated by a statistical machine translation system into German. We then investigated how the summarization affected readability and MT.

2.1 Summarizer

We used the extractive summarizer COGSUM (Jönsson et al., 2010). It is based on the vector space model random indexing (RI). Each sentence in a document is scored by RI, and then ranked using a weighted PageRank algorithm. A summary is then created by choosing the top sentences, which supposedly are the most important to the document. The length of the summaries can be varied; in the current study we use 50% summaries, i.e., extract half the sentences from each text.

We used a version of COGSUM with an external corpus, the Brown Corpus, for training the vector space, which has been shown to be more stable than training the vector space on the document to be summarized (Smith and Jönsson, 2011b).

2.2 Readability measures

Readability is a complex notion, that is related both to texts and to individual readers and their skills. However, many automatic readability measures have previously been proposed based on, often superficial, text properties. In the current study we chose to work with two such measures: the Flesch reading ease formula (Flesch, 1951), which has commonly been used for English, and LIX (Björnsson, 1968), which has traditionally been used for Swedish. Flesch is based on average sentence length (ASL) and average syllable length, and LIX is based on ASL and the proportion of long words. In addition, we report ASL on its own.

2.3 Machine translation system

We used a factored phrase-based statistical MT system (Koehn and Hoang, 2007) with treatment of compound words, to translate from English to German, as described in Holmqvist et al. (2011). The system was trained on 1.7M sentences from Europarl, 136K sentences of News text, and an additional 17M sentences of monolingual news data. The system is evaluated using the Bleu metric (Papineni et al., 2002).

2.4 Data set

The experiment was performed on news stories from the development data of the WMT12 workshop.1 From this data, we extracted all coherent news stories with at least 20 sentences, resulting in 236 news stories, with a total of 7602 sentences. The stories were translated by the SMT system, and an analysis was performed on both the full set of sentences, and on sentences that occurred in the summaries, compared to sentences not in the summaries. As we found that the summarization affected the average sentence length in these sets, we also used test sets balanced for sentence length, by extracting an equal amount of sentences in all three conditions, with lengths between 13–30 words per sentence, in total 1556 sentences.

1http://statmt.org/wmt12
3. Result

In this section we first present the results of summarization on readability, followed by the effect of summarization and readability indicators on machine translation.

3.1 Readability for summaries

To analyze the effect on readability of the summarization we computed readability measures for each of the 236 stories on the full test set (All), the summarized sentences (Sum), and the sentences not chosen for summarization (Nosum). A summary of the results is shown in Table 1. For Flesch, a high score means more readable, whereas a low score is better for LIX and ASL. All differences from the original texts are significant with \( p < 0.01 \). On all three measures the summaries are less readable than the original texts. We also found that there is a strong significant correlation of -0.91 between Flesch and LIX, something that we believe has not been shown before, since these metrics have tended to be used for different languages.

<table>
<thead>
<tr>
<th></th>
<th>Flesch↑</th>
<th>LIX↓</th>
<th>ASL↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>44.1</td>
<td>49.6</td>
<td>21.4</td>
</tr>
<tr>
<td>Sum</td>
<td>39.6</td>
<td>54.1</td>
<td>27.0</td>
</tr>
<tr>
<td>Nosum</td>
<td>47.9</td>
<td>45.6</td>
<td>15.6</td>
</tr>
</tbody>
</table>

Table 1: Readability before and after summarization

3.2 Effects on MTranslatability

Table 2 shows the Bleu scores for the three data partitions, both full and normalized in length. Note that the test sets are different, for the full sets they even have different size, and are thus not directly comparable. The scores, however, indicate that the summaries, which were the least readable, have lower scores than the non-summaries, even when length is compensated for.

<table>
<thead>
<tr>
<th></th>
<th>Full sets</th>
<th>length normalized</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>15.4</td>
<td>15.0</td>
</tr>
<tr>
<td>Sum</td>
<td>15.1</td>
<td>15.2</td>
</tr>
<tr>
<td>Nosum</td>
<td>15.7</td>
<td>15.7</td>
</tr>
</tbody>
</table>

Table 2: Bleu scores

To allow a further analysis of the interaction between SMT, readability and summarization, we performed a multiple linear regression analysis on blocks of 50 sentences for each test set. We could find no significant correlations between Bleu and readability measures or summarization set. We could, however, confirm the correlation between Flesch and LIX, which in this configuration was -0.92. Figure 1 illustrates this for the full test sets.

4. Conclusion

In this study summarization resulted in longer and less readable sentences, which is contrary to previous research (Smith and Jönsson, 2011a). The language used and summarizer configuration were different in these studies, however. We plan to conduct further research on the effect of different summarization methods in future work.

None of the readability factors or summarization sets contributed to any significant differences in MT quality in this study. The overall Bleu scores, though, at least seemed to indicate that there is some relation to readability, since the nosum test set had both the highest Bleu scores, and the best readability. The current study has several limitations, though, such as the use of automatic metrics. For instance, even though Bleu has been shown to have some correlation with human judgment, it might not be the best option for this type of study, as in CE, where human judgments have been more useful than metrics. We believe that these issues are interesting, and that further analysis of the interplay between summarization, MTranslatability, and readability could be valuable.

5. References


