Overview of the Hebrew Dataset MultiLing 2013

Tal Baumel, Raphael Cohen, Michael Elhadad, Sagit Fried, Avi Hayoun, Yael Netzer
Computer Science Dept.
Ben-Gurion University in the Negev
<lastname>@cs.bgu.ac.il

This document describes the process of preparing the dataset for MultiLing 2013 in Hebrew: translation of source texts from English, and the summarization for the translated texts, by the Ben Gurion University Natural Language Processing team.

1. Translation Process

Four people participated in the translation and the summarization of the dataset of the 50 news articles: three graduate students, one a native English speaker with fluent Hebrew and the other two with Hebrew as a mother tongue and very good English skills. The process was supervised by a professional translator with a doctoral degree with experience in translation and scientific editing.

The average times to read an article was 2.5 minutes (std. dev 1.2min), the average translation time was 30 minutes (std. dev 15min), and the average proofing time was 18.5min (std. dev 10.5min).

2. Translation Methodology

We tested two translation methodologies by different translators. In some of the cases, translation was aided with Google Translate[1], while in other cases, translation was performed from scratch.

In the cases where texts were first translated using Google Translate, the translator reviewed the text and edited changes according to her judgment. Relying on the time that was reported for the proofreading of each translation, we could tell that texts that were translated using this method, required longer periods of proofreading (and sometimes more time was required to proofread than to translate). This is most likely because once the automatic translation was available, the human translator was biased by the automatic outcome, remaining anchored' to the given text with reduced criticism and creativity.

Translating the text manually, aided with online or offline dictionaries, Wikipedia and news site on the subject that was translated, showed better quality as analysis of time shows, where the ratio between the time needed to proofread was less than half.

In addition, we found, that in most cases the time that the translation took for the first texts of a given subject (for each article cluster), tends to be significantly longer than the subsequent articles in the same cluster. This reflects the 'learning phase' experienced by the translators who approached each cluster, getting to know the vocabulary of each subject.

3. Clusters Topics

The text collection includes five clusters of ten articles each. Some of the topics were very familiar to the Hebrew-speaking readers, and some subjects were less familiar or relevant. The Iranian Nuclear issue is very common in the local news and terminology is well known. Moreover, it was possible to track the articles from the news as they were published in Hebrew news websites at that time; this was important for the usage of actual and correct news-wise terminology. The hardest batch to translate was on the Paralympics championship, which had no publicity in Hebrew, and the terminology of winter sports is culturally foreign to native Hebrew speakers.

4. Special Issues in Hebrew

A couple of issues have surfaced during the translation and should be noted. Many words in Hebrew have a foreign transliterated usage and an original Hebrew word as well. For instance, the Latin word *Atomic* is very common in Hebrew and, therefore, it will be equally acceptable to use it in the Hebrew form, אטומי / *'atomi'* but also the Hebrew word גרעיני (*'gar'ini' / nuclear*). Traditional Hebrew News Agen-

[1] http://translate.google.com/#

cies have for many years adopted an editorial line which strongly encourages using original Hebrew words whenever possible. In recent years, however, this approach is relaxed, and both registers are equally accepted. We have tried to use a 'common notion' in all texts using the way terms are written in Wikipedia as the voice of majority. In most cases, this meant using many transliterations.

Another issue in Hebrew concerns the orthography variations of plene *vs*. deficient spelling. Since Hebrew can be written with or without vocalization, words may be written with variations. For instance, the vocalized version of the word 'air' is אֲוִיר ('*avir*') while the non-vocalized version is אוויר ('*avvir*'). The rules of spelling related to these variations are complicated and are not common knowledge. Even educated people write words with high variability, and in many cases, usage is skewed by the rules embedded in the Microsoft Word editor. We did not make any specific effort to enforce standard spelling in the dataset.

5. Summarization Process

Each cluster of articles was summarized by three persons, and each summary was proof-read by the other summarizers. Most of the summarizers read the texts before summarization, while translating or proofreading them, and, therefore, the time that was required to read all texts was reduced.

The time spent reading and summarizing was extremely different for each of the three summarizers, reflecting widely different summarization strategies, as indicated in the following table (average times over 5 clusters):

Summarizer	**Reading time**	**Summarizatio n**
A	43 min	49 min
B	22 min	84 min
C	35 min	62 min

The trend indicates that investing more time up front reading the clusters pays off later in summarization time.

The instructions did not explicitly recommend abstractive vs. extractive summarization. Two summarizers applied abstractive methods, one tended to use mostly extractive (C). The extractive method did not take markedly less time than the abstractive one. In the evaluation, the extractive summary was found markedly less fluent.

As the best technique to summarize efficiently, all summarizers found that ordering the texts by date of publication was the best way to conduct the summaries in the most fluent manner.

However, it was not completely a linear process, since it was often found that general information, which should be located at the beginning of the summary as background information, appeared in a later text. In such cases, summarizers changed their usual strategy and consciously moved information from a later text to the beginning of the summary. This was felt as a distinct deviation – as the dominant strategy was to keep track of the story told across the chronology of the cluster, and to only add new and important information to the summary that was collected so far.

The most difficult subject to summarize was the set on Paralympic winter sports championship which was a collection of anecdotal descriptions which were not necessarily a developing or a sequential story and had no natural coherence as a cluster.