

TIDES Multilingual Summarization Evaluation (MSE) 2005

Instructions to Annotators

1 Introduction

TIDES Multilingual Summarization Evaluation (MSE) targets the development of summarization technology. Imagine that you are a news analyst, or one who needs to synthesize or absorb a large amount of information in as little time as possible. Summarization technology will assist such a person by automatically-generating summaries (or “blurbs”) about a topically related cluster of documents.

The goal of MSE 2005 is to first select topic “clusters,” then to create 100-word topic summaries.

1.1 Corpus

The corpus for MSE is comprised of the English and Arabic portions of the TDT-4 corpus, which is a large collection stemming from multiple sources (newswire, broadcast news transcripts). The corpus covers an epoch from October 2000 to January 2001.

LDC requested machine-translations of the entire TDT-4 Arabic corpus from the Information Sciences Institute (ISI) at the University of Southern California. Summaries are created using the translated data.

1.2 Terms

Topic – for the purposes of MSE, a topic is a very narrowly defined event, and only that event. For example, an earthquake could be a MSE topic, but only reports that discuss the actual earthquake would be “on topic”; reports of a state of emergency being declared or of reconstruction efforts would *not* be on-topic.

Cluster – a cluster is a group of topically-related documents. For MSE clusters will contain an even distribution of Arabic and English documents.

2 Topic Selection

The first task LDC must complete is topic selection, which will involve examining a series of topic clusters that were automatically generated by Columbia University’s clustering system. The topic clusters are grouped by “super cluster”, then by “sub cluster”. Our task is to read through the documents in each sub cluster and evaluate the cohesion to the topic. A “topic” is a group of documents that discuss a single, narrowly-defined event. LDC annotators will develop 25 topics by sifting through the pre-generated clusters and identifying the most topically relevant ones.

3 Topic Summarization

Once the 25 test topics are identified, LDC will create 100-word summaries of those topics, and will develop four independent summaries for each, resulting in 100 total summaries.

3.1 Annotation tool

LDC uses an in-house summarization annotation tool, which in one window displays to the annotator a concatenated list of all documents in each topic cluster. A short heading or description precedes the documents and serves as a topic reference. The second window of the tool displays a free-text entry box where annotators write the 100-word summaries. In the lower margin of the interface a word-count mechanism keeps track of how many words are in each summary.

3.2 Summary instructions

After accessing the summarization annotation tool, select a topic from the topic list. Read through all documents – both translated and original English – and synthesize the important information about the topic event. Taking notes on paper or in another text editor is strongly advised; this will keep the important facts organized while you continue to read. Once you have read all of the documents in the cluster, compose a 100-word resume of that topic, capturing the important people, places, and details surrounding the topic event. As you write, watch how many words you have, and edit as necessary.

3.2.1 Guidelines

Correct spelling and grammar is required. Write out numerals under the number ten (10). Do not abbreviate days of the week or months of the year. Rely on the English documents for the correct spelling of proper nouns; the translated documents often err on the spelling of people and place names.

When you have completed a topic summary, run a spell check. If it is over or fewer than 100 words, edit it until the summary reaches 100. Sometimes this means returning to the summary for that topic, which you are permitted to do.

4 Quality Control

LDC will exact a few quality control measures; namely, one or more passes over each summary. Each “pass” by lead annotators or managers will check a summary’s orthography, topic relevance, word count (100), and double-check proper nouns.