A Crude Cut at Query Expansion

Phil Rennert
StreamSage, Inc.
phil.rennert@streamsage.com

I had planned to approach the Novelty task by using machine learning methods to fill templates. However, the paucity of training topics (three; I invalidated 379 due to the large inter-assessor disagreement) and the sparseness of relevant-but-not-new sentences would not support this approach.

To do what I could with the training data, I experimented with query expansion. For search terms, I took all the nouns in the title; and added two categories of expansion:

> 1) Multi-word units. StreamSage has a corpus of New York Times news stories, from which we've extracted all multi-word units (words which occur together in sequence) meeting certain grammatical and frequency criteria. I selected all MWUs containing a title noun, and added all other nouns in that MWU as search terms.
>
> 2) Noun phrases in the description (desc, not desc1). Any nouns occurring in the same noun phrase with a title noun were added as search terms.

Then I returned any sentence containing a search term as relevant.

To determine which sentences were new, I used a pure length criterion. The shortest sentences were returned relevant-but-not-new, in the same proportion as in the training data (where 5-10% of the relevant sentences were assessed as not-new).

I hope to do more next year, with 50 training topics available.