A machine learning approach for QA and Novelty Tracks: NTT system description

Hideto Kazawa, Tsutomu Hirao, Hideki Isozaki and Eisaku Maeda
NTT Communication Science Laboratories
Nippon Telegraph and Telephone Corporation
{kazawa,hirao,isozaki,maeda}@cslab.kecl.ntt.co.jp

1 A unified approach to QA and Novelty Tasks

In one sense, the goals of QA and Novelty tasks are the same: extracting small document parts which are relevant to users' queries. Additionally, the unit of extraction is almost always fixed in both tasks. For QA, an answer is a noun phrase in most cases, and for Novelty, a sentence is recognized as the basic information unit.

This observation leads us to the following unified approach to both QA and Novelty tasks: first identify information units in documents, then judge whether each unit is relevant to the query. This two step approach is amenable to machine learning methods because each step can be cast as a classification problem. For example, noun phrase identification can be achieved by classifying each word into the start/middle/end/exterior of a noun phrase; sentence identification by classifying whether each period marks the of a sentence. Additionally, relevance judgment can be regarded as the classification of a pair of query and an information unit into a relevant-pair or non-relevant-pair.

In QA and Novelty Tracks at TREC 2002, we studied the feasibility of this two step approach, using Support Vector Machines as the learning algorithm of the classifiers. Since many studies on identifying information units have already been reported, we concentrate on the relevance judgment step in QA and Novelty tasks in this paper.

2 Question Answering Track

Because of limited time, we applied our machine learning approach only to questions concerning dates and quantities; the other questions were processed in the same way as reported in [1]. Hereafter, we limit ourselves to the questions about dates and quantities.

2.1 Answer Candidate Identification

For date/quantity questions, answers are likely base noun phrases (base NPs), including date or number. Thus, we extracted base NPs including date/number expressions as information units (answer candidates). However, since identifying such base NPs is arguably rather easy work, we constructed an answer candidate identification module based on a Naive-Bayes classifier instead of SVMs [1].

2.2 Relevant Candidate Selection

To train SVMs on the relevance jugdement of candidates, we created a training dataset that consists of pairs of date/quantity questions and their answers.

As positive example pairs, we used the pairs of the past QA Track questions that ask dates/quantities, and the answers collected from the past judgment files using answer patterns. In addition, we randomly selected incorrect answer candidates from the judgement files, then combined them with the questions to obtain negative example pairs.

Each question-candidate pair is converted into a feature vector. The features consists of the following two types of features.

Keyword densities

Keyword density in 5/10/20-word Hanning windows centered on the candidate.

Combined features

All combinations of question and candidate features. The question features consist of:

- Wh-word in the question,
- Headword of wh-phrase in the question and its WordNet category,
- Keywords (nouns, verbs) in the question and its WordNet category.

Here, wh-words are who, when, where, and what, whereas wh-phrases are phrases including wh-words.

The candidate features consist of:

- Headword of the candidate and its WordNet category,
- Binary indicator of whether the candidate includes number/month/day expressions,
- The number of digits in the candidates.

3 Novelty Track

3.1 Relevant Sentence Extraction

We made a training data set that consists of query-sentence pairs whose relevance was judged by us. Our data includes 21 queries chosen from TREC topics, which are not used in the novelty track, and 4044 sentences chosen from past TREC results.

To apply SVM, we transformed each query-sentence pair into a feature vector. The features consist of:

[1] In general, SVMs show higher performance than Naive-Bayes classifiers. However, we prefer Naive-Bayes in this case because training Naive-Bayes classifiers is very fast.

- Sentence position normalized by the document length.
- Sentence length normalized by the longest sentence length in the same document.
- The sum of the weights of the sentence vector.
- Keyword density in the sentence. The keywords are terms in the description section of the query and the title section of the document.
- Cosine between the headline term vector and the sentence term vector.
- Cosine between the query term vector and the sentence term vector.
- Cosine between the query term vector and the document term vector.

Here, term vectors are commonly-used tfidf vectors. The inverse document frequency (idf) is calculated using all the TIPSTER document sets.

For the TREC 2002 Novelty Track, we trained SVMs with the quadratic kernel, $(x \cdot x' + 1)^2$, and the Gaussian kernel, $\exp(-a|x - x'|^2)$.

3.2 "New" Sentence Selection

For the TREC 2002 Novelty Track, it is not sufficient to merely judge the relevance of each information unit (sentence); it is required that only "new" sentences be reported at the end.

To select "new" sentences from relevant sentences, we used Marginal Relevance (MR) as the selection criteria. Originally, MR was proposed as a measure of "information" increased by the addition of a new document to the documents already selected[2].

$$\mathrm{MR}_{\mathcal{D}}(d) = \lambda \mathrm{rel}(d) - (1 - \lambda) \max_{d' \in \mathcal{D}} \mathrm{sim}(d, d'). \tag{1}$$

Here, d is the document whose contribution to information increase is to be measured, and $\mathcal{D}$ is the set of documents already selected. The function $\mathrm{rel}(d)$ is the relevance of d to the query, and $\mathrm{sim}(d, d')$ is the "similarity" between d and d'.

To apply MR to a new sentence selection in the Novelty task, we modified MR as follows.

- Sentences are regarded as (very) short "documents".
- For the relevance measure of sentences, we use the value of a discriminant function which is used in the relevant sentence extraction step.
- For the similarity measure, we use the number of the common words between the sentences divided by the average number of words in both sentences.

In the new sentence selection step, we repeat the selection of the sentence with the largest MR and add it to the selected sentence set until the largest MR is less than 0.2 [2]

[2]We set $\lambda = 0.7$.

	Quadratic	Gaussian
P*R(rel)	0.0694	0.0664
P*R(new)	0.0545	0.0477

Table 1: Results of Novelty Track

3.3 Results

We submitted two runs for Novelty Tracks: quadratic kernel SVMs were used in one run and Gaussian kernel SVMs in the other.

Table 1 shows "precision multiplied by recall" values of our submissions. Quadratic kernel SVMs perform slightly better than Gaussian kernel ones.

References

[1] H. Kazawa, H. Isozaki and E. Maeda, "NTT Question Answering System in TREC 2001," Proc. of TREC 2001, pp.415–422. (2001)

[2] J. Carbonell and J. Goldstein, "The Use of MMR, Diversity-Based Reranking for Reordering Document and Producing Summaries," Proc. of SIGIR-98, pp.335–336. (1998)