Learning to Rank Definitions to Generate Quizzes for Interactive Information Presentation

Ryuichiro Higashinaka and **Kohji Dohsaka** and **Hideki Isozaki**
NTT Communication Science Laboratories, NTT Corporation
2-4, Hikaridai, Seika-cho, Kyoto 619-0237, Japan
{rh,dohsaka,isozaki}@cslab.kecl.ntt.co.jp

Abstract

This paper proposes the idea of ranking definitions of a person (a set of biographical facts) to automatically generate "Who is this?" quizzes. The definitions are ordered according to how difficult they make it to name the person. Such ranking would enable users to interactively learn about a person through dialogue with a system with improved understanding and lasting motivation, which is useful for educational systems. In our approach, we train a ranker that learns from data the appropriate ranking of definitions based on features that encode the importance of keywords in a definition as well as its content. Experimental results show that our approach is significantly better in ranking definitions than baselines that use conventional information retrieval measures such as tf*idf and pointwise mutual information (PMI).

1 Introduction

Appropriate ranking of sentences is important, as noted in sentence ordering tasks (Lapata, 2003), in effectively delivering content. Whether the task is to convey news texts or definitions, the objective is to make it easier for users to understand the content. However, just conveying it in an encyclopedia-like or temporal order may not be the best solution, considering that interaction between a system and a user improves understanding (Sugiyama et al., 1999) and that the cognitive load in receiving information is believed to correlate with memory fixation (Craik and Lockhart, 1972).

In this paper, we discuss the idea of ranking definitions as a way to present people's biographical information to users, and propose ranking definitions to automatically generate a "Who is this?" quiz. Here, we use the term 'definitions of a person' to mean a short series of biographical facts (See Fig. 1). The definitions are ordered according to how difficult they make it to name the person. The ranking also enables users to easily come up with answer candidates. The definitions are presented to users one by one as hints until users give the correct name (See Fig. 2). Although the interaction would take time, we could expect improved understanding of people's biographical information by users through their deliberation and the long lasting motivation afforded by the entertaining nature of quizzes, which is important in tutorial tasks (Baylor and Ryu, 2003).

Previous work on definition ranking has used measures such as tf*idf (Xu et al., 2004) or ranking models trained to encode the likelihood of a definition being good (Xu et al., 2005). However, such measures/models may not be suitable for quiz-style ranking. For example, a definition having a strong co-occurrence with a person may not be an easy hint when it is about a very minor detail. Certain descriptions, such as a person's birthplace, would have to come early so that users can easily start guessing who the person is. In our approach, we train a ranker that learns from data the appropriate ranking of definitions. Note that we only focus on the ranking of definitions and not on the interaction with users in this paper. We also assume that the definitions to be ranked are given.

Section 2 describes the task of ranking definitions, and Section 3 describes our approach. Section 4 describes our collection of ranking data and the ranking model training using the ranking support vector machine (SVM), and Section 5 presents the evaluation results. Section 6 summarizes and mentions future work.

2 Ranking Definitions for Quizzes

Figure 1 shows a list of definitions of Natsume Soseki, a famous Japanese novelist, in their original ranking at the encyclopedic website goo (*http://dictionary.goo.ne.jp/*) and in the quiz-style ranking we aim to achieve. Such a ranking would realize a dialogue like that in Fig. 2. At the end of the dialogue, the user would be able to associate the person and the definitions better, and it is expected that some new facts could be learned about that person.

Proceedings of the ACL 2007 Demo and Poster Sessions, pages 117–120,

Original Ranking:

1. Novelist and scholar of British literature.
2. Real name: Kinnosuke.
3. Born in Ushigome, Edo.
4. Graduated from the University of Tokyo.
5. Master of early-modern literature along with Mori Ogai.
6. After the success of “I Am a Cat”, quit all teaching jobs and joined Asahi Shimbun.
7. Published masterpieces in Asahi Shimbun.
8. Familiar with Haiku, Chinese poetry, and calligraphy.
9. Works include “Botchan”, “Sanshiro”, etc.

⇓

Quiz-style Ranking:

1. Graduated from the University of Tokyo.
2. Born in Ushigome, Edo.
3. Novelist and scholar of British literature.
4. Familiar with Haiku, Chinese poetry, and calligraphy.
5. Published masterpieces in Asahi Shimbun.
6. Real name: Kinnosuke.
7. Master of early-modern literature along with Mori Ogai.
8. After the success of “I Am a Cat”, quit all teaching jobs and joined Asahi Shimbun.
9. Works include “Botchan”, “Sanshiro”, etc.

Figure 1: List of definitions of Natsume Soseki, a famous Japanese novelist, in their original ranking in the encyclopedia and in the quiz-style ranking. The definitions were translated by the authors.

Ranking definitions is closely related to definitional question answering and sentence ordering in multi-document summarization. In definitional question answering, measures related to information retrieval (IR), such as tf*idf or pointwise mutual information (PMI), have been used to rank sentences or information nuggets (Xu et al., 2004; Sun et al., 2005). Such measures are used under the assumption that outstanding/co-occurring keywords about a definiendum characterize that definiendum. However, this assumption may not be appropriate in quiz-style ranking; most content words in the definitions are already important in the IR sense, and strong co-occurrence may not guarantee high ranks for hints to be presented later because the hint can be too specific. An approach to creating a ranking model of definitions in a supervised manner using machine learning techniques has been reported (Xu et al., 2005). However, the model is only used to distinguish definitions from non-definitions on the basis of features related mainly to linguistic styles.

In multi-document summarization, the focus has been mainly on creating cohesive texts. (Lapata, 2003) uses the probability of words in adjacent sentences as constraints to maximize the coherence of all sentence-pairs in texts. Although we acknowledge that having cohesive definitions is important, since we are not creating a single text and the dialogue that we aim to achieve would involve frequent user/system interaction (Fig. 2), we do not deal with the coherence of definitions in this paper.

S1	Who is this? First hint: Graduated from the University of Tokyo.
U1	Yoshida Shigeru?
S2	No, not even close! Second hint: Born in Ushigome, Edo.
U2	I don’t know.
S3	OK. Third hint: Novelist and scholar of British literature.
U3	Murakami Haruki?
S4	Close! Fourth hint: Familiar with Haiku, Chinese poetry, and calligraphy.
U4	Mori Ogai?
S5	Very close! Fifth hint: Published masterpieces in Asahi Shimbun.
U5	Natsume Soseki?
S6	That’s right!

Figure 2: Example dialogue based on the quiz-style ranking of definitions. S stands for a system utterance and U for a user utterance.

3 Approach

Since it is difficult to know in advance what characteristics are important for quiz-style ranking, we learn the appropriate ranking of definitions from data. The approach is the same as that of (Xu et al., 2005) in that we adopt a machine learning approach for definition ranking, but is different in that what is learned is a quiz-style ranking of sentences that are already known to be good definitions.

First, we collect ranking data. For this purpose, we turn to existing encyclopedias for concise biographies. Then, we annotate the ranking. Secondly, we devise a set of features for a definition. Since the existence of keywords that have high scores in IR-related measures may suggest easy hints, we incorporate the scores of IR-related measures as features (*IR-related features*).

Certain words tend to appear before or after others in a biographical document to convey particular information about people (e.g., words describing occupations at the beginning; those describing works at the end, etc.) Therefore, we use word positions within the biography of the person in question as features (*positional features*). Biographies can be found in online resources, such as biography.com (*http://www.biography.com/*) and Wikipedia. In addition, to focus on the particular content of the definition, we use bag-of-words (BOW) features, together with semantic features (e.g., semantic categories in Nihongo Goi-Taikei (Ikehara et al., 1997) or word senses in WordNet) to complement the sparseness of BOW features. We describe the features we created in Section 4.2. Finally, we create a ranking model using a preference learning algorithm, such as the ranking SVM (Joachims, 2002), which learns ranking by reducing the pairwise ranking error.

4 Experiment

4.1 Data Collection

We collected biographies (in Japanese) from the goo encyclopedia. We first mined Wikipedia to calculate the PageRank™ of people using the hyper-link structure. After sorting them in descending order by the PageRank score, we extracted the top-150 people for whom we could find an entry in the goo encyclopedia. Then, 11 annotators annotated rankings for each of the 150 people individually. The annotators were instructed to rank the definitions assuming that they were creating a "who is this?" quiz; i.e., to place the definition that is the most characteristic of the person in question at the end. The mean of the Kendall's coefficients of concordance for the 150 people was sufficiently high at 0.76 with a standard deviation of 0.13. Finally, taking the means of ranks given to each definition, we merged the individual rankings to create the reference rankings. An example of a reference ranking is the bottom one in Fig. 1. There are 958 definition sentences in all, with each person having approximately 6–7 definitions.

4.2 Deriving Features

We derived our IR-related features based on Mainichi newspaper articles (1991–2004) and Wikipedia articles. We used these two different sources to take into account the difference in the importance of terms depending on the text. We also used sentences, sections (for Wikipedia articles only) and documents as units to calculate document frequency, which resulted in the creation of five frequency tables: (i) Mainichi-Document, (ii) Mainichi-Sentence, (iii) Wikipedia-Document, (iv) Wikipedia-Section, and (v) Wikipedia-Sentence.

Using the five frequency tables, we calculated, for each content word (nouns, verbs, adjectives, and unknown words) in the definition, (1) frequency (the number of documents where the word is found), (2) relative frequency (frequency divided by the maximum number of documents), (3) co-occurrence frequency (the number of documents where both the word and the person's name are found), (4) relative co-occurrence frequency, and (5) PMI. Then, we took the minimum, maximum, and mean values of (1)–(5) for all content words in the definition as features, deriving 75 ($5 \times 5 \times 3$) features. Then, using the Wikipedia article (called an *entry*) for the person in question, we calculated (1)–(4) within the entry, and calculated tf*idf scores of words in the definition using the term frequency in the entry. Again, by taking the minimum, maximum, and mean values of (1)–(4) and tf*idf, we yielded 15 (5×3) features, for a total of 90 (75 + 15) IR-related features.

Positional features were derived also using the Wikipedia entry. For each word in the definition, we calculated (a) the number of times the word appears in the entry, (b) the minimum position of the word in the entry, (c) its maximum position, (d) its mean position, and (e) the standard deviation of the positions. Note that positions are either ordinal or relative; i.e., the relative position is calculated by dividing the ordinal position by the total number of words in the entry. Then, we took the minimum, maximum, and mean values of (a)–(e) for all content words in the definition as features, deriving 30 (5×2 (ordinal or relative positions) $\times$ 3) features.

For the BOW features, we first parsed all our definitions with CaboCha (a Japanese morphological/dependency parser, *http://chasen.org/~taku/software/cabocha/*) and extracted all content words to make binary features representing the existence of each content word. There are 2,156 BOW features in our data.

As for the semantic features, we used the semantic categories in Nihongo Goi-Taikei. Since there are 2,715 semantic categories, we created 2,715 features representing the existence of each semantic category in the definition. Semantic categories were assigned to words in the definition by a morphological analyzer that comes with ALT/J-E, a Japanese-English machine translation system (Ikehara et al., 1991).

In total, we have 4,991 features to represent each definition. We calculated all feature values for all definitions in our data to be used for the learning.

4.3 Training Ranking Models

Using the reference ranking data, we trained a ranking model using the ranking SVM (Joachims, 2002) (with a linear kernel) that minimizes the pairwise ranking error among the definitions of each person.

5 Evaluation

To evaluate the performance of the ranking model, following (Xu et al., 2004; Sun et al., 2005), we compared it with baselines that use only the scores of IR-related and positional features for ranking, i.e., sorting. Table 1 shows the performance of the ranking model (by the leave-one-out method, predicting the ranking of definitions of a person by other peo-

Rank	Description	Ranking Error
1	**Proposed ranking model**	**0.185**
2	Wikipedia-Sentence-PMI-max	0.299
3	Wikipedia-Section-PMI-max	0.309
4	Wikipedia-Document-PMI-max	0.312
5	Mainichi-Sentence-PMI-max	0.318
6	Mainichi-Document-PMI-max	0.325
7	Mainichi-Sentence-relative-co-occurrence-max	0.338
8	Wikipedia-Entry-ordinal-Min-max	0.338
9	Wikipedia-Sentence-relative-co-occurrence-max	0.339
10	Wikipedia-Entry-relative-Min-max	0.340
11	Wikipedia-Entry-ordinal-Mean-mean	0.342

Table 1: Performance of the proposed ranking model and that of 10 best-performing baselines.

ple's rankings) and that of the 10 best-performing baselines. The ranking error is pairwise ranking error; i.e., the rate of misordered pairs. A descriptive name is given for each baseline. For example, Wikipedia-Sentence-PMI-max means that we used the maximum PMI values of content words in the definition calculated from Wikipedia, with sentence as the unit for obtaining frequencies.

Our ranking model outperforms all of the baselines. McNemar's test showed that the difference between the proposed model and the best-performing baseline is significant ($p<0.00001$). The results also show that PMI is more effective in quiz-style ranking than any other measure. The fact that max is important probably means that the mere existence of a word that has a high PMI score is enough to raise the ranking of a hint. It is also interesting that Wikipedia gives better ranking, which is probably because people's names and related keywords are close to each other in such descriptive texts.

Analyzing the ranking model trained by the ranking SVM allows us to calculate the weights given to the features (Hirao et al., 2002). Table 2 shows the top-10 features in weights in absolute figures when all samples were used for training. It can be seen that high PMI values and words/semantic categories related to government or creation lead to easy hints, whereas semantic categories, such as birth and others (corresponding to the person in 'a person from Tokyo'), lead to early hints. This supports our intuitive notion that birthplaces should be presented early for users to start thinking about a person.

6 Summary and Future Work

This paper proposed ranking definitions of a person to automatically generate a "Who is this?" quiz. Using reference ranking data that we created manually, we trained a ranking model using a ranking SVM based on features that encode the importance of keywords in a definition as well as its content.

Rank	Feature Name	Weight
1	Wikipedia-Sentence-PMI-max	0.723
2	SemCat:33 (others/someone)	-0.559
3	SemCat:186 (creator)	0.485
4	BOW:*bakufu* (feudal government)	0.451
5	SemCat:163 (sovereign/ruler/monarch)	0.422
6	Wikipedia-Document-PMI-max	0.409
7	SemCat:2391 (birth)	-0.404
8	Wikipedia-Section-PMI-max	0.402
9	SemCat:2595 (unit; e.g., numeral classifier)	0.374
10	SemCat:2606 (plural; e.g., plural form)	-0.368

Table 2: Weights of features learned for ranking definitions by the ranking SVM. SemCat denotes it is a semantic-category feature with its semantic category ID followed by the description of the category in parentheses. BOW denotes a BOW feature.

Experimental results show that our ranking model significantly outperforms baselines that use single IR-related and positional measures for ranking. We are currently in the process of building a dialogue system that uses the quiz-style ranking for definition presentation. We are planning to examine how the different rankings affect the understanding and motivation of users.

References

Amy Baylor and Jeeheon Ryu. 2003. Does the presence of image and animation enhance pedagogical agent persona? *Journal of Educational Computing Research*, 28(4):373–395.

Fergus I. M. Craik and Robert S. Lockhart. 1972. Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior*, 11:671–684.

Tsutomu Hirao, Hideki Isozaki, Eisaku Maeda, and Yuji Matsumoto. 2002. Extracting important sentences with support vector machines. In *Proc. 19th COLING*, pages 342–348.

Satoru Ikehara, Satoshi Shirai, Akio Yokoo, and Hiromi Nakaiwa. 1991. Toward an MT system without pre-editing –Effects of new methods in ALT-J/E–. In *Proc. Third Machine Translation Summit: MT Summit III*, pages 101–106.

Satoru Ikehara, Masahiro Miyazaki, Satoshi Shirai, Akio Yokoo, Hiromi Nakaiwa, Kentaro Ogura, Yoshifumi Ooyama, and Yoshihiko Hayashi. 1997. *Goi-Taikei – A Japanese Lexicon*. Iwanami Shoten.

Thorsten Joachims. 2002. Optimizing search engines using clickthrough data. In *Proc. KDD*, pages 133–142.

Mirella Lapata. 2003. Probabilistic text structuring: Experiments with sentence ordering. In *Proc. 41st ACL*, pages 545–552.

Akira Sugiyama, Kohji Dohsaka, and Takeshi Kawabata. 1999. A method for conveying the contents of written texts by spoken dialogue. In *Proc. PACLING*, pages 54–66.

Renxu Sun, Jing Jiang, Yee Fan Tan, Hang Cui, Tat-Seng Chua, and Min-Yen Kan. 2005. Using syntactic and semantic relation analysis in question answering. In *Proc. TREC*.

Jinxi Xu, Ralph Weischedel, and Ana Licuanan. 2004. Evaluation of an extraction-based approach to answering definitional questions. In *Proc. SIGIR*, pages 418–424.

Jun Xu, Yunbo Cao, Hang Li, and Min Zhao. 2005. Ranking definitions with supervised learning methods. In *Proc. WWW*, pages 811–819.