# Query classification by using named entity recognition systems and clue keywords

Masaharu Yoshioka
Graduate School of Information Science and Technology, Hokkaido University
N14 W9, Kita-ku, Sapporo-shi
Hokkaido Japan
yoshioka@ist.hokudai.ac.jp

## ABSTRACT

Query classification is a subtask of 1CLICK task for selecting appropriate strategy to generate output text. In this paper, we propose to use named entity recognition tools and clue keywords (occupation name list and location type name list) to identify query types.

## Team Name

HUKB

## Subtasks

Query classification task (Japanese)

## Keywords

Named Entity Recognition, Wikipedia

## 1. INTRODUCTION

1CLICK aims to satisfy the user with a single textual output instead of a ranked list of URLs [1]. Query classification task is a subtask of 1CLICK to predict the query type (ARTIST, ACTOR, POLITICIAN, ATHLETE, FACILITY, GEO, DEFINITION and QA). This information is useful to select appropriate strategy to generate output text (e.g., strategy to generate text for QA may be different from one for ATHLETE).

In this paper, we propose to use named entity recognition tools and clue keywords (occupation name list and location type name list) to identify query types.

## 2. QUERY CLASSIFICATION SYSTEM

### 2.1 Approach for query classification

The basic idea for this query classification process is listed as below.

1. Query normalization
   In the collection, there are several queries that name of a named entity are separated by using white space. In order to normalize the style of representing a named entity in queries, we check the connectivity of the query terms to identify the named entity boundary.

   For example, “加藤 康弘 所属クラブ” (Kato Yasuhiro Club: J-0035), is normalized as “加藤康弘 所属クラブ”, since “加藤 康弘” is a person name that are separated by a white space.

2. Usage of named entity recognition tool
   In order to identify the basic category of the query terms, named entity recognition tools are used for extracting PERSON, ORGANIZATION and LOCATION keyword.

   We assume queries with person name are classified into one of the four person types (ARTIST, ACTOR, POLITICIAN, and ATHLETE), and ones with ORGANIZATION and LOCATION are classified into GEO or FACILITY.

3. Usage of POS tagger
   POS tagger is used for extracting interrogative words and verbs. Query with interrogative word are classified into QA. Query with verb may be classified into QA.

4. Extraction of definition sentences
   We assume query for DEFINITION may have define statement in the relevant documents.

5. Prediction of occupation for a person
   In order to classify the person into four categories (ARTIST, ACTOR, POLITICIAN, and ATHLETE), we made a list of occupation for each category. In addition, we predict occupation of the person from the relevant documents.

6. Location type keywords
   GEO is a query to find entities with geographic constraints. Therefore, this type of query has geographic constraints keyword (mostly represented by LOCATION or ORGANIZATION) and location type keyword (e.g, restaurant, hospital, and theater). A list of location type keywords is used for identify query for GEO.

### 2.2 Query classification procedure

Figure 1 shows a basic procedure to classify the query. Followings are explanation of each step.

### *Query normalization*

Query normalization is a process to detect word boundary of one or more named entities in a query. Followings are assumption for this process.

- When two or more keywords are part of a name, most of the relevant documents contain all combined keywords as a phrase.
- Web search engine may retrieve enough number of relevant documents as a search result.

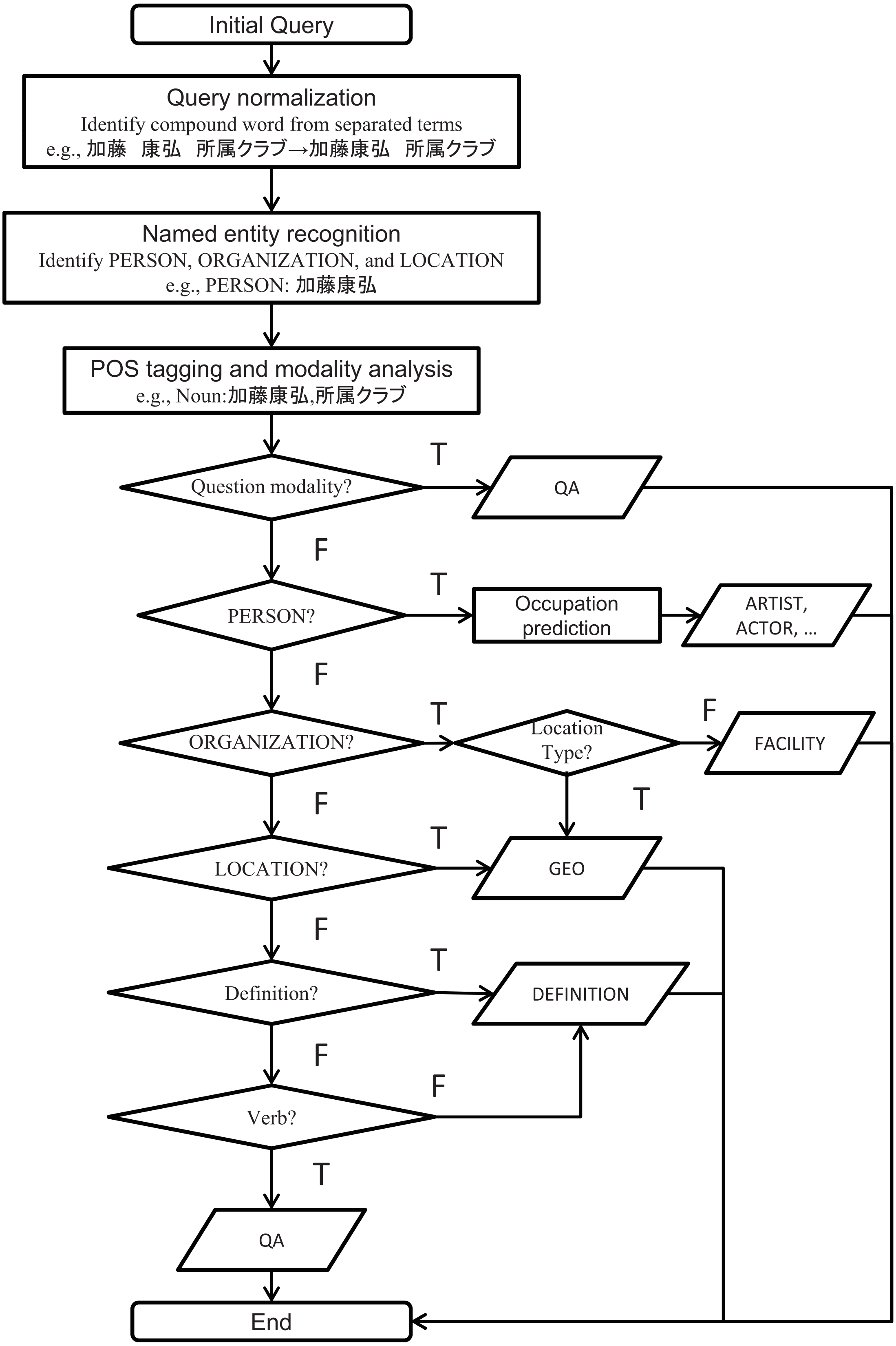

**Figure 1: Flowchart of query classification**

Based on these assumptions, we conduct following procedures for every query term boundary to check whether it is connected as a part of named entity or not.

1. Calculation of separated terms and combined phrase frequency
   Since we would like to check whether a combined phrase is a part of named entity or not, it is necessary to discuss how these terms are combined for representing a named entity. There are varieties of method for make a phrase. For example, Japanese person name may be combined with a white space and foreign name may be combined by special character such as “・” or “=”. Therefore we count term frequency ($TF_{phrase}$) of phrase connected with white space, “・”, “=” or no space by using regular expression from given documents (ORCL run: all relevant documents, MANDATORY run: top 10 documents). Term frequency of each terms ($TF_1, TF_2$) are also calculated.

2. Comparison between $TF_{phrase}$ and $TF_1, TF_2$.
   Official name (e.g., person name described as a combination of family name and given name) may be abbreviated by using a part of the name (e.g., family name). Therefore frequency of such abbreviated term, which may have higher frequency than others, is not meaningful to decide the connectivity of terms. In addition, maximal number of $TF_{phrase}$ is bounded by $min(TF_1, TF_2)$.

   So, following formula is used for check whether it is connected as a part of named entity or not.

   $TF_{phrase} \leq \alpha \times min(TF_1, TF_2)$

   $\alpha = 0.3$ is used throughout this experiment.

### Named Entity recognition

In order to have higher recall for named entity recognition, we use KNP [1] and CaboCha [2][2].

KNP uses a large named entity dictionary based on Wikipedia information and extract hypernym by using Wikipedia category. However, since KNP selects only one hypernym for the named entity, it may miss to select appropriate occupation for a person with multiple ones. Therefore, we can not rely on the output of KNP for extracting occupation.

CaboCha is also used for having higher recall for named entity recognition.

### POS tagging and modality analysis

JUMAN is used for POS tagging and KNP is used for modality analysis. Question modality is a strong clue to identify QA type query. In addition, ARTIST, ACTOR, POLITICIAN, ATHLETE, FACILITY, and GEO are categories for representing named entity and are represented by nouns. On the contrary, DEFINITION and QA may include verbs. Therefore, existence of verb by using POS tagging for identify QA or DEFINITION type query.

### Definition (explanation) sentences extraction

A definition (explanation) sentence is a sentence that gives definition (explanation) to a given word. Typical examples of the sentences are the first sentences of Wikipedia entries. These sentences may contains clue keywords to identify the type of query.

For example, first sentence of “畑 恵”(Kei Hata) in the Wikipedia is “畑 恵（はた・けい、女性、本名の姓:船田（ふなだ）、旧姓・畑。1962年（昭和37年）2月15日 - ）は、日本の政治家、フリーアナウンサー。”(“Kei Hata(HATA KEI, ..., 1962.2.15-) is a Japanese politician and freelance announcer)”). From this sentence, we can extract occupation of “畑恵” as politician and freelance announcer (POLITICIAN is a correct answer of a sample query set distributed by organizer).

In addition, since DEFINITION type query may have such sentences, it is also a good clue to identify DEFINITION type.

Extraction of definition (explanation) sentence is conducted by using simple pattern matching as follows.

1. Deletion of additional information described in parenthesis
   Especially in Wikipedia, explanation sentence may have additional information such as date of birth, and how to pronounce by using parenthesis. Such parenthesis part are removed for conducting simple pattern matching.

2. Simple pattern matching
   For the query term Q, sentences contain “Q とは” are extracted as definition sentences.

### Occupation prediction

In order to predict occupation for named entities, occupation list is constructed by using Wikipedia category. In the Wikipedia, “職業別の人物” (People by occupation) category is a top category for occupation categories. From these subcategories, we extract 162 occupation names and 10 suffix terms (e.g., “医” and “士” ) and classify these occupations into 4 categories (ARTIST, ACTOR, POLITICIAN, and ATHLETE).

Procedure for occupation prediction is as follows.

1. Extraction occupation from definition (explanation) sentences
   We conduct pattern matching to extract candidate occupations from these sentences. For occupation extraction of query term Q, sentences contain “Q は” are also extracted as definition (explanation) sentences.

2. Extraction occupation from other sentences
   When no occupation is extracted at the first step, we conduct pattern matching to extract candidate occupations from other sentences.

3. Evaluation of candidate occupations
   As we discussed before, there is a possibility to have two or more candidate occupations for a named entity. In such a case, we evaluate appropriateness of the occupation (O) of a named entity (P) by counting out sentences that match “O の P” (P, O or O P) patterns from relevant documents.

   When there is no sentences that match these patterns for candidate occupation from definition (explanation) sentences, extraction occupation from other sentences is conducted for finding other candidates.

   For example, in the case of “畑恵”(Kei Hata), number of sentences that contains “政治家の畑恵” and “フ

[1] http://nlp.ist.i.kyoto-u.ac.jp/index.php?KNP
[2] http://code.google.com/p/cabocha/

リーアナウンサーの畑恵" are counted. When there is no sentences for these patterns, candidates from other sentences such as "参議院議員" (member of the House of Councillors) are extracted and number of sentences that contains "参議院議員の畑恵" are counted.

Finally, all counts of occupations are merged into counts of corresponding category and select most frequent categories as an appropriate category.

However, POLITICIAN is a special case for this category. For example, "畑恵" (Kei Hata) is ex-member of the House of Councillors, but she still work as POLITICIAN. "谷亮子" (Ryoko Tani) is famous gold medalist, but she is also a member of the House of Councillors. Since we would like to categorize them into POLITICIAN, we decide POLITICIAN is a special category and who have occupation of POLITICIAN category is selected as POLITICIAN.

### *Location type string matching*

GEO type query is one that combines keyword type of named entity and geographical constraints. Since name of a ORGANIZATION or LOCATION can be used as geographical constraints, queries that have ORGANIZATION keywords can be categorized FACILITY or GEO. In order to distinguish these two categories, it is necessary to check type of additional keywords to the facility keyword.

It is not so easy to make the list of keywords for additional information, we decide to make a list of location type name (e.g., restraints, shops, hospitals) for this purpose. We construct this list by using location type name for location search service (Genre code API of Yahoo Open Local Platform [3]). ORGANIZATION keyword(s) with location type name and LOCATION keyword(s) are selected as GEO. ORGANIZATION keyword(s) without location type name is selected as FACILITY.

### *Definition type identification*

Finally, queries without category selection is checked for DEFINITION or QA. Query keywords that have definition (explanation) sentences are selected as DEFINITION. Query keywords that have verb are selected as QA. At last, queries without category selection is selected as DEFINITION.

## 3. EXPERIMENT

Based on the procedure discussed in previous section, we submit two runs that use all retrieved web page as relevant documents (MANDATORY) and a list of relevant documents only (ORCL).

Table shows precision of each run.

Followings are typical errors for our system.

- Named entity recognition error (MANDATORY: 11 ORCL:11)
  Named entity recognition tools fails to identify many facilities (6 out of 15 for FACILITY query) and CaboCha tends to misclassify unknown terms (that can not be parsed by using pre-defined diction nary) as ORGANIZATION.

- Occupation prediction error (MANDATORY: 2 ORCL:5)
  Patterns for occupation prediction ("O の P") is widely used in news article style documents, but those documents are excluded from ORCL. Due to lack of these documents quality of ORCL is worse than MANDATORY. Errors of MANDATORY related to the heuristics of identify POLITICIAN. It is better to reconsider the heuristics related to identify POLITICIAN.

| | MANDATORY | ORCL |
|---|---|---|
| ARTIST | 10/10 | 10/10 |
| ACTOR | 8/10 | 4/10 |
| POLITICIAN | 10/10 | 9/10 |
| ATHLETE | 8/10 | 9/10 |
| FACILITY | 9/15 | 9/15 |
| GEO | 15/15 | 15/15 |
| DEFINITION | 9/15 | 9/15 |
| QA | 11/15 | 11/15 |
| TOTAL | 80/100 | 78/100 |

**Table 1: Number of Correct answer for all query class**

- POS and Modality analysis error (MANDATORY:4 ORCL:4)
  Several QA queries are not explained in question modality. It is necessary to have another method to estimate its type.

- Definition (explanation) sentence extraction error (MANDATORY: 2 ORCL:2)
  We assume DEFINITION queries may have corresponding definition sentences "Q とは", but there are several queries about artifact (e.g., "ザ・ウォール" (The Wall: J-0077)) and character of comic (e.g., "犬夜叉" (Inuyasha: J-0072). Since those queries have different style for explanation, our system fails to classify such data.

- Other (MANDATORY: 1 ORCL:1)
  Query "感謝祭 カナダ" (Thanksgiving Day Canada: J-0081) contains location name and is similar to "GEO" and our system classify this query as GEO.

From this analysis, we confirm the method for job prediction generally works well when there are enough numbers of news article style documents. However, it is necessary to have a mechanism to check the quality of named entity recognition system (e.g., implementing facility type prediction that is similar to the occupation prediction proposed in this paper).

## 4. CONCLUSION

In this paper, we propose a query classification method by using named entity recognition tools and clue keywords (occupation name list and location type name list). We confirm our strategy works well for most of the queries. However, failure analysis shows that it is necessary to have a mechanism for handling inappropriate results of named entity recognition system.

## 5. REFERENCES

[1] M. P. Kato, M. Ekstrand-Abueg, V. Pavlu, T. Sakai, T. Yamamoto, and M. Iwata. Overview of the

[3] http://developer.yahoo.co.jp/webapi/map/openlocalplatform/v1/genreCode.html

NTCIR-10 1CLICK-2 task. In *Proceedings of the 10th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Quesiton Answering, And Cross-Lingual Information Access*, 2013. (to appear).

[2] T. Kudo and Y. Matsumoto. Japanese dependency analysis using cascaded chunking. In *CoNLL 2002: Proceedings of the 6th Conference on Natural Language Learning 2002 (COLING 2002 Post-Conference Workshops)*, pages 63–69, 2002.