Automatic Extraction of Named Entity Related Relations for Searching

Nguyen Thanh Tri
School of Information Science
t-thanh@jaist.ac.jp

1. The aim of research

Named entities (NEs) play an important role in many Natural Language Processing applications including semantic search. Discovering the NE-related relations may be beneficial to these applications. Our study proposes a to extract relations of NEs in the form of ⟨*ne, category, related-to, object*⟩ quadruples which describe that the named entity *ne* ISA $category$, and the $category$ IS-RELATED-TO $object$. We extended the Person Category Extraction (PCE) algorithm to extract these tuples, and experiments on Wall Street Journal (WSJ) corpus give promising results.

2. The approach

Text documents often contain valuable relations of entities. For example, in the sentence taken from the WSJ corpus:

There's a generally more positive attitude toward the economy, said Bette Raptapoulos, analyst for Prudential-Bache Securities Inc., ... (1)

there are relations: "Bette Raptapoulos" is-a analyst, and analyst for "Prudential-Bache Securities Inc." Such relations may be beneficial in many NLP applications, such as for answering *Who* and *List* questions, e.g., "Who is Bette Raptapoulos?" or "Give me the list of analyst for Prudential-Bache Securities Inc."

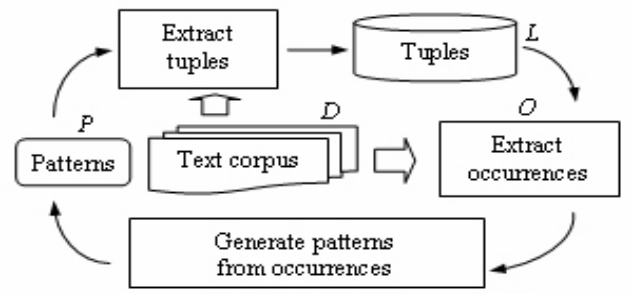

Figure 1. PCE's model

Person Category Extraction:

The purpose of NECE is to extract ⟨*named_entity, category*⟩ tuples, in which $category$ is the fine-grained categories of $named\_entity$, so the set of named entity classes can be expanded by automatically extracting from texts. When we extract the tuple ⟨"Bette Raptapoulos", "analyst"⟩ from (1), we only have information: "Bette Raptapoulos" ISA "analyst". If we can extract the relation: "analyst" **for** "Prudential-Bache Securities Inc.", we will have complete information about "Bette Raptapoulos". The PCE algorithm is depicted in Figure 1. Starting with two seed patterns, PCE extracted ⟨*person, category*⟩ tuples. The extracted tuples were used to extract the occurrences of ⟨*person, category*⟩ tuples in texts for generating new patterns. Again, new patterns were used to extract new tuples. The process terminated when no more patterns were produced. A pattern is defined as a 4-tuple: (*order, person_slot, middle, category_pattern*), where *order* indicates the occurrence order of person and category in a sentence; *person_slot* is a slot which will be replaced with a person named entity; *middle* is the string surrounded by person and category; *category_pattern* is defined as: *category_pattern*:=*noun_phrase1 (and noun_phrase2)*? where $noun\_phrase_i$ is a regular expression that matches a noun phrase with added POS tags. In order to extract new ⟨*person, category*⟩ tuples, for every sentence s, from a pattern, and for each person NE *named_entity* in s, we construct a regular expression:

* *named_entity middle category_pattern* *

If s matches the above regular expression, the ⟨*person, category*⟩ tuple is extracted. For improving the performance of the algorithm, at each match if the *category* is not valid (i.e., the $category$ is not a type of person) the match is ignored.

An occurrence of a person, category tuple is defined as a 4-tuple: ⟨*order, person, middle, category*⟩, where middle is a string surrounded by person and category. An occurrence of a (person, category) tuple is extracted if a sentence s matches the regular expression: * *person middle category* * or * *category middle person* * After having been extracted, occurrences are used to generate new patterns. A $middle$ of an occurrence is not necessarily reliable,

some constraints were proposed to retain reliable ones: Repetition of a middle (repetition(middle)) is the number of times the middle appears between the person and category of ⟨*person, category*⟩ tuples of same person. Diversity of a middle ($diversity(middle)$) is the number of times the middle appears between the person and category of ⟨*person, category*⟩ tuples of different persons. A middle that has $repetition(middle) > threshold_R$ seems reliable and is kept. A pattern seems specific if it is generated based on tuples of a person, so only middles that have $diversity(middle) > threshold_D$ are kept to make the generated patterns general (Condition 1). If a middle contains a verb phrase, the verb phrase should express the relation person ISA category (Condition 2).

Named-Entity-Related Relations Extraction:

We extend the PCE to extract ⟨*named_entity, category, related-to, object*⟩ quadruples describing the relations: *named_entity* ISA *category* (or ISA relations for short), and category *related-to* object (or *related-to* relations for short). From our observations, the related-to relations can be expressed in the following ways:

a) The $category$ and $object$ are linked by a preposition: "*category preposition object*", e.g., "analyst for Prudential-Bache Securities Inc."

b) The *category* and *object* are connected by a possessive apostrophe, e.g., "Semi-Tech's chief executive officer". This can be interpreted as "category of object", e.g., "chief executive officer of Semi-Tech".

c) The *object* and *named_entity* are linked by a preposition, e.g., "... said economist David Littmann of Manufacturers National Bank ...", from which an expected quadruple is ⟨"David Littmann", "economist", "of", "Manufacturers National Bank"⟩.

d) The *object* is embedded in *category*, e.g., "IBM president". This can also be interpreted as "*category* of *object*", e.g., "president of IBM".

e) The *related-to* relation is implicitly expressed, e.g., "Mr. Baird, who heads the Manhattan U.S. attorney's securities-fraud unit, denied the quote ...", from which an expected quadruple is ⟨"Baird", "head", "of", "securities-fraud unit"⟩.

Since case e) does not have fixed expressions, we do not treat such cases. In case d), because object is already embedded in category, we do not need to extract the object. For cases a), b) and c), we construct regular expressions to extract the *object* and *related-to*.

Utilization of Named-Entity-Related Relations for Semantic Search:

The extracted (*named_entity, category, related-to, object*) quadruples are valuable for NLP applications. In this section, we use them for answering some types of questions. If *named_entity* in a quadruple is a person, the quadruple helps answer the query: "Who is *named_entity*?", e.g., "Who is Bette Raptapoulos?" If *named_entity* is of another type, such as *organization* or *location*, the quadruple helps answer the query: "What is *named_entity*?", e.g., "What is IBM?" For answering the question, we just search for quadruples having the same *named_entity* as that of the question. If a quadruple is found, then the answer is: "*named_entity* is a(n) *category related-to object*".

The extracted quadruples also help answer *list* questions, e.g., "Give me the list of analyst for Prudential-Bache Securities Inc." The general form of this question type is "Give me the list of *category* [*related-to object*]", where the part in square brackets is optional. For answering this question type, we search for the list *L* of quadruples having the same *category*, [*related-to, and object*] as those of the question. The answer is the list of *named_entity* of quadruples in L.

3 The progress of the search

Our work was published in the "Proceedings of the 21st Pacific Asia Conference on Language, Information and Computation" (PACLIC21).

4 Future direction

There are still other named-entity-related relations in documents, and discovering such relations may help answer other question types. In the future, we intend to study to extract other relations in documents for searching.

5 Publication

1. **T. T. Nguyen** and A. Shimazu, "Automatic Extraction of the Fine Category of Person Named Entities from Text Corpora", *IEICE Transactions on Information and Systems, Special section on Knowledge, Information and Creativity Support System*, Vol. E90-D, No. 10, pp. 1542-1549, 2007.
2. **T. T. Nguyen** and A. Shimazu, "Acquisition of Named-Entity-Related Relations for Searching", *Proceedings of the 21st Pacific Asia Conference on Language, Information and Computation*, (PACLIC21), pp. 349-357, 2007.
3. **T. T. Nguyen**, L. M. Nguyen, and A. Shimazu, "Using Semi-supervised Learning for Question Classification", *Journal of Natural Language Processing*, Vol. 15, No. 1, 2008. (to appear)