Exploiting Dependency Context Gazetteers for Named Entity Recognition

Han-Cheol Cho[1]　Naoaki Okazaki[2,3]　Kentaro Inui[2]

[1]The University of Tokyo (hccho@is.s.u-tokyo.ac.jp)

[2]Tohoku University ({okazaki, inui}@ecei.tohoku.ac.jp)

[3]Japan Science and Technology Agency (JST)

Abstract

Modern named entity recognition (NER) systems mostly employ a supervised machine learning approach that heavily depends on local contexts. While NER systems based on local contexts provide strong baseline performance, results of recent research have demonstrated that non-local contexts can further improve the performance of these systems. In this paper, we propose the use of a context gazetteer as a novel resource for improving NER with non-local information. A context gazetteer is a list of syntactic dependency contexts with which entity names co-occur. We build a context gazetteer from a large encyclopedic database because manually annotated data are often too few to extract rich and sophisticated context patterns. Moreover, each context is assigned with a confidence value to reflect its reliability. In experiments, we create a context gazetteer of gene names and apply it to a biomedical NER task. High confidence context patterns appear in various forms: some are similar to predicate–argument structures whereas some are in unexpected forms. The experimental results show that the context gazetteer improves both precision and recall over strong baseline models.

1　Introduction

High performance of supervised NER systems require a set of features that are well designed to distinguish entity mentions from others. It is well known that local features, which can obtained from a small linear context window (local context hereinafter), contribute to production of strong baseline models [6]. For example, presuming that we shall determine the label of the underlined word “associated” in Fig. 1, the neighboring and current words such as “major”, “plastid-lipid”, “associated”, “protein” and “is” within the local context [-2,2] are useful as word uni-gram features. However, recent studies [3, 6] have demonstrated that non-local contexts can provide useful information that local contexts can not supply. In Fig. 1, for instance, direct and indirect head-words of the word “associated” such as “protein”, “encoding”, “gene”, and “expressed” are very informative because all of them are semantically related to genes.

In this paper, we propose to use a context gazetteer, which is a list of contexts that co-occur with entity names, for incorporating new sentence level non-local features into NER models. A context gazetteer consists of dependency paths of variable lengths to capture syntactically meaningful contexts more than traditional local contexts. Moreover, confidence values are assigned to contexts to reflect how much they are likely to co-occur with specific entity types. Unlike previous studies [1] using only manually annotated data, we build a context gazetteer from a huge amount of precisely labeled data.

In experiment, we build a context gazetteer of gene names and apply it to a biomedical NER task. It is particularly interesting that top-ranked entries in the context gazetteer appear in various forms. As expected, there are many predicate–argument style contexts with domain specific predicates such as “express”, “inhibit” and “promote.” However, they also frequently appear in unexpected forms such as abbreviation, apposition and conjunction dependencies. These contexts can be interpreted as fragments of domain knowledge that appear in stereotypical syntactic structures in texts. When the context gazetteer is applied, both precision and recall improves.

The remainder of this paper is organized as follows. Section 2 describes the proposed method for creating a context gazetteer. In the next section, we build a context gazetteer of gene names from the EntrezGene database, and apply it to the BioCreative 2 gene name recognition task [7]. The usefulness of a context gazetteer is demonstrated experimentally. We also analyze what kinds of context patterns are mined and how they affect NER models. Section 4 summarizes the contributions of this work, and explains remaining future work.

2　Building a Gazetteer

A context gazetteer is a weighted list of dependency paths (hereinafter, contexts) of variable length that co-occur with target entity names. Figure 2 portrays an exemplary context of length 3; a word X is likely to be an entity word, which is a part of a target entity name, surrounded by the context consisting of the head word *expression*, a dependent *cells* and a grand-dependent *cancer* with the corresponding dependencies *prep_ of*, *prep_ in* and *nn*. This context can help to recognize the underlined gene name in a sentence, “The *expression* of FasL in gastric *cancer*

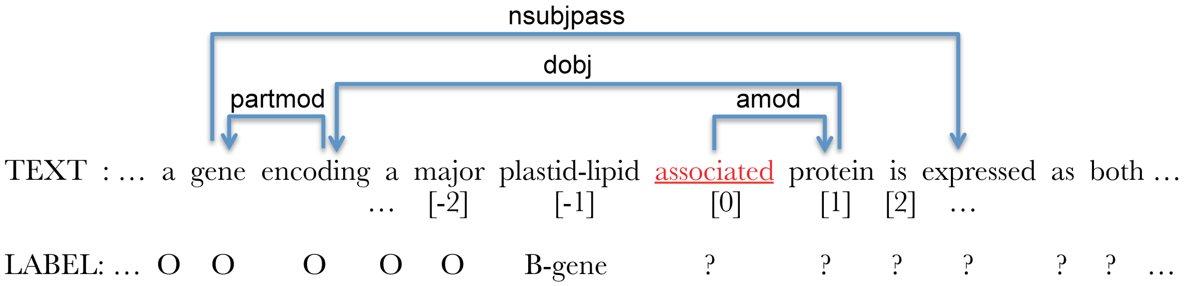
nsubjpass
dobj
partmod
amod
TEXT : … a gene encoding a major plastid-lipid associated protein is expressed as both …
… [-2] [-1] [0] [1] [2] …
LABEL: … O O O O O B-gene ? ? ? ? ? ? …

Figure 1: The local context window [-2,2] is shown under the text, whereas the non-local context window is shown with directed arrows. “plastid-lipid associated protein” is a gene name. The definition of dependency labels are explained in the Stanford dependency manual.

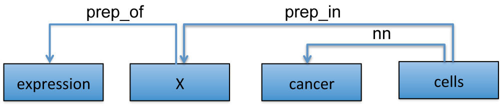

Figure 2: An example context of the length 3.

cells and of Fas in apoptotic TIL was also detected in vivo.”

A good context gazetteer should have 1) rich contexts for high coverage and 2) reliable contexts for high precision. For the first requirement, we extract contexts from a large amount of automatically labeled data rather than a small manually annotated data. To satisfy the second requirement, we assign confidence values to the extracted contexts. Figure 3 shows a flow of building a context gazetteer satisfying these requirements. The remainder of this section explains each step in detail, and we apply this process to a real word problem in the next section.

Step.1 Automatic Text Labeling: A straightforward approach to obtain a large amount of labeled data is to label in-domain texts using a number of target entity names. This approach labels every occurrence of each entity name in the texts with the corresponding entity type such as person, organization and location. However, the labeled data is inevitably very noisy because most entity names are ambiguous in the absence of contexts. For example, it is hard to tell whether the word “Inception” is a movie title or a general noun without contexts. Moreover, entity names may have multiple entity types. For instance, person names can constitute the names of companies (e.g., Ford Motor Company), diseases (e.g. Alzheimer disease), places (e.g., Washington, D.C) and so on.

To solve this problem, we adopted an approach in the previous study [9]. In short, we use an encyclopedic database consisting of target entity names and their descriptions because entity names in their description are mostly unambiguous.

Step.2 Context Extraction: The labeled texts are then parsed and the dependency paths (contexts) involving entity words are extracted. At this step, the number of extracted contexts is very large. We filtered out the contexts that have no content words (nouns, verbs and adjectives) except for an entity word because these contexts are often too general.

Step.3 Context Normalization: For each context, an entity word is substituted with a placeholder X as shown in Fig. 2. To increase the coverage of a context gazetteer, it is necessary to perform normalization. Because normalization is often often domain-specific techniques, we will explain them in Sec. 3.1 while actually building a context gazetteer of gene names.

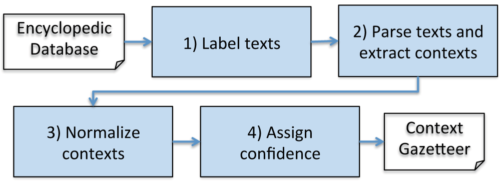

Figure 3: Procedure for building a context gazetteer.

Step.4 Confidence Assignment: Contexts are often ambiguous even if they frequently appear with specific entity types. We solve this problem by assigning a confidence to each context for every entity type. Assuming that text data D is automatically annotated with entity names of T different entity types[1], the confidence is defined as the conditional probability of an entity type t given a context c as in

$$\text{confidence}(t|c) = p(t|c) = \frac{C(t,c)}{C(c)} = \frac{\sum_{e_t \in D} C(e_t,c)}{C(c)}. \tag{1}$$

In this equation, $C(c)$ is the frequency of the context c in D, $C(t,c)$ is the frequency that the context c and the entity type t co-occur in D. $C(t,c)$ can be calculated by $\sum_{e_t \in D} C(e_t,c)$ because the occurrence of the entity type t is equal to the occurrence of entity words e_t belonging to the entity type t.

3 Evaluation

To demonstrate the usefulness of a context gazetteer, we apply the proposed method to the BioCreative 2 gene mention recognition task [7].

3.1 Data Preparation

Context Gazetteer. For building a context gazetteer, we use gene names (including synonyms)

[1] The set T includes non-entity type O too.

Conf.	Pattern
1.0	nsubj(globin, X)
1.0	prep_between(interaction, X) ∧ conj_and(X, C-Jun)
1.0	prep_for(screened, mutations) ∧ prep_of(mutations, gene) ∧ nn(gene, X)
0.91	prep_of(secretion, X) ∧ amod(X, inhibitory)
0.81	nsubj(repressor, X)
0.78	prep_as(X, mediator)
0.65	dobj(express, X)
0.55	nsubjpass(known, function) ∧ prep_of(function, X)

Table 1: Examples of high confidence context patterns. Conf. stands for confidence.

of the EntrezGene database[2] and the abstracts of MEDLINE[3] articles linked from the EntrezGene. Automatically labeled abstracts (358,049 in total) are parsed using the Stanford CoreNLP tool[4]. Then, dependency paths (contexts) that involve entity words are extracted. The maximum length of contexts is set to be 5 experimentally. For domain-specific normalization, continuous numbers and symbols of the words are converted into a representative number (0) and symbol (under-bar), respectively. Contexts appearing less than 10 times are removed because estimated confidence can be unreliable.

Several extracted contexts having high confidence are presented in Table 1. At the beginning of this study, we expected to obtain contexts conveying domain specific knowledge, especially in predicate–argument structure (PAS). For example, the second, fourth and the seventh contexts are all in the form of PAS using nominal and verbal predicates. The second context indicates that X is likely to be a gene if it appears has an interaction with *C-jun* as in "... interaction between X and C-Jun." However, we also found unexpected but interesting contexts too. First, many contexts capture factual knowledge. The first and fifth contexts are the simplest ones meaning that X is likely to be a gene if it is a *globin* or a *repressor*. The sixth context means X is likely to be a gene if it acts as a *mediator*. Second, some contexts represent procedural information. The third context, for instance, indicates that there is a screening process for analyzing mutations of a gene. Lastly, the eighth context, seemingly uninformative at first glance, means that discovering the function of a gene is a common task as in "The exact function of IP-30 is not yet known, but it may play a role ..."

Entity Gazetteer. Entity gazetteers are one of the most important resources for NER. Four entity gazetteers are compiled from the EntrezGene, the Universal Protein Resource (UniProt), the Unified Medical Language System (UMLS) and the Open Biological and Biomedical Ontologies (OBO). For improving the coverage of these gazetteers, continuous numbers and symbols of the entity names are normalized into a representative number and symbol (0 for numbers and under-bar for symbols), and all alphabet characters are lower-cased.

Syntactic Analysis. We also used the GENIA tagger [8] to perform lemmatization, POS-tagging and chunking; however, we disabled the NER module and did not use its results.

Model	E.G.	C.G.	Prec.	Rec.	F1
Base1	none	none	87.99	81.71	84.73
Prop1	none	Entrez	88.06	81.42	84.61
Base2	Entrez	none	88.54	82.17	85.24
Prop2	Entrez	Entrez	88.66	82.99	85.73
Base3	All	none	89.06	82.78	85.81
Prop3	All	Entrez	89.32	83.46	86.29

Table 2: Performance evaluation using entity and context gazetteers. E.G.: Entity Gazetteer, C.G.: Context Gazetteer, Entrez: EntrezGene, All: EntrezGene, UniProt, UMLS, and OBO.

Features. Baseline models use features that are common in most NER systems such as unigrams and bigrams of tokens, lemmas, POS-tags, the combination of lemmas and POS-tags, and entity gazetteers within a local context window [-2,2]. In addition, the chunk type of a current token and the last token of a current chunk are used. To relieve unknown word problem, we also exploit orthographic features of a current token [4]. Proposed models use the context gazetteer in addition to these local features. A token surrounded by context(s) of the context gazetteer is tagged with context gazetteer class label. The confidence of a context is quantized at every 0.1 step. For example, if a token is surrounded by two contexts with the confidence 0.31 and 0.56, then we assign two labels to the token, "ContextGaz_EntrezGene_3" and "ContextGaz_EntrezGene_6", where the confidence is rounded up.

Machine Learning. For machine learning, we use the CRFsuite [5], which implements first-order linear-chain Conditional Random Fields. The regularization parameter (C) is optimized using the first 90% of the original training data as training data and the rest 10% as the development data. Fifteen *C* values (0.03125, 0.0625, 0.125, 0.25, 0.5, 0.75, 1, 2, 3, 4, 5, 6, 8, 10, and 16) are tested ad the best performing one is chosen.

3.2 Experiment Results

Table 2 shows an experiment result that evaluates the effect of a context gazetteer in combination with various entity gazetteers. We prepared three baseline models. The *Base1* model uses no gazetteer at all, whereas the *Base2* model uses the entity gazetteer compiled from the EntrezGene and the *Base3* model

[2] EntrezGene is National Center for Biotechnology Information (NCBI) ’ s database for gene-specific information.

[3] MEDLINE is the U.S. National Library of Medicine’sÂő (NLM) premier bibliographic database.

[4] `http://nlp.stanford.edu/software/corenlp.shtml`

uses all four entity gazetteers. Three proposed models corresponding to these baseline models employ the context gazetteer built from the EntrezGene. For the *Prop2* and *Prop3* models, the context gazetteer improves the performance in both precision and recall. In addition, the context gazetteer increases recall notably that is harder to improve than precision [2] due to the asymmetric label distribution where one class label, O, dominates all other classes.

However, the performance of the *Prop1* model drops slightly compared to the *Base1* model.

3.3 Result Analysis

We manually compared about 20% of the output of the *Base3* and *Prop3* models to see how the context gazetteer features affect the tagging results.

There are 32 gene names correctly recognized by the *Prop3* but not by the *Base3*. In all of these cases, one or more context gazetteer features are triggered. The following list shows several examples in which the *Prop3* recognized the under-barred gene names and the *Base3* recognized the italicized gene names.

1. One major transcript encodes MEQ, a *339-amino-acid bZIP protein* which is homologous to the Jun/Fos family of transcription factors.

2. The association of I-92 with *p92*, *p84*, *p75*, *p73*, *p69*, and *p57* was completely reversible after treatment with the detergent deoxycholate (DOC).

3. The exact function of IP-30 is not yet known, but it may play a role in *gamma-interferon* mediated immune reactions.

In the first example, two context gazetteer features "dobj(encode, X)" and "appos(X, protein)" are triggered for the gene name "MEQ." The second feature is a strong evidence of X being a gene name because a word X is in apposition with the word protein. In the second example, "I-92" has a context gazetteer feature "prep_of(association, X) ∧ prep_with(X, p0)" meaning that X is likely to be a part of gene name if it is associated with the gene name "p0" where 0 represents any number. Contexts of these kinds are the fragments of domain specific knowledge and usually have high confidence. In the last example, the gene name "IP-30" has a context gazetteer feature "nsubjpass(known, function) ∧ prep_of(function, X)" with the confidence 0.54. This feature shows a stereotypical expression often used in the introduction section of an article.

However, 15 gene names are not recognized by the *Prop3* model whereas they are correctly recognized by the *Base3* model. For 3 cases, no context gazetteer features are triggered. We suspect that the coverage of the context gazetteer may not be high enough because we use words (not stems or lemmas) in the contexts. For the other 12 cases, context gazetteer features are fired but not recognized by the *Prop3* model.

4 Conclusion and Future Work

This paper proposed the use of a context gazetteer as a new non-local feature for NER. Compared to the feature aggregation methods [6], the proposed method can be easily applied to streaming data such as tweets and pre-processed data with sentence selection where recognizing document (or discourse) boundaries is difficult. We also described how to induce a rich and sophisticated context gazetteer from automatically annotated data using an encyclopedic database. The proposed method is applied to a biomedical NER task and its usefulness is demonstrated in combination with entity gazetteers.

However, we also found that the coverage of the context gazetteer is not high enough. For this research, we used words and their dependencies as contexts. However, these contexts sometimes include uninformative words in the middle of contexts. If it is possible to generalize the contexts by replacing these unimportant words with POS-tags or wildcards, then the coverage of the context gazetteer can be enhanced.

Acknowledgement: This research was partly supported by JST, PRESTO. This research was partly supported by JSPS KAKENHI Grant Numbers 23240018 and 23700159

References

[1] Jenny Finkel, Shipra Dingare, Huy Nguyen, Malvina Nissim, Christopher Manning, and Gail Sinclair. Exploiting context for biomedical entity recognition: from syntax to the web. In *Proceedings of the International Joint Workshop on NLPBA*, pp. 88–91, 2004.

[2] Nanda Kambhatla. Minority vote: at-least-n voting improves recall for extracting relations. In *Proceedings of COLING-ACL*, pp. 460–466, 2006.

[3] Vijay Krishnan and Christopher D. Manning. An effective two-stage model for exploiting non-local dependencies in named entity recognition. In *Proceedings of COLING-ACL*, pp. 1121–1128, 2006.

[4] Ki-Joong Lee, Young-Sook Hwang, Seonho Kim, and Hae-Chang Rim. Biomedical named entity recognition using two-phase model based on svms. *Journal of Biomedical Informatics*, Vol. 37, No. 6, pp. 436–447, December 2004.

[5] Naoaki Okazaki. Crfsuite: A fast implementation of conditional random fields (crfs), 2007.

[6] Lev Ratinov and Dan Roth. Design challenges and misconceptions in named entity recognition. In *Proceedings of the Thirteenth Conference on CoNLL*, pp. 147–155, 2009.

[7] Larry Smith, Lorraine Tanabe, Rie Ando, Cheng-Ju Kuo, I-Fang Chung, Chun-Nan Hsu, Yu-Shi Lin, Roman Klinger, Christoph Friedrich, Kuzman Ganchev, Manabu Torii, Hongfang Liu, Barry Haddow, Craig Struble, Richard Povinelli, Andreas Vlachos, William Baumgartner, Lawrence Hunter, Bob Carpenter, Richard Tsai, Hong-Jie Dai, Feng Liu, Yifei Chen, Chengjie Sun, Sophia Katrenko, Pieter Adriaans, Christian Blaschke, Rafael Torres, Mariana Neves, Preslav Nakov, Anna Divoli, Manuel Mana-Lopez, Jacinto Mata, and W John Wilbur. Overview of biocreative ii gene mention recognition. *Genome Biology*, Vol. 9, No. Suppl 2, p. S2, 2008.

[8] Yoshimasa Tsuruoka and Jun'ichi Tsujii. Bidirectional inference with the easiest-first strategy for tagging sequence data. In *Proceedings of the Conference on HLT-EMNLP*, pp. 467–474, 2005.

[9] Yu Usami, Han-Cheol Cho, Naoaki Okazaki, and Jun'ichi Tsujii. Automatic acquisition of huge training data for bio-medical named entity recognition. In *Proceedings of BioNLP 2011 Workshop*, pp. 65–73, 2011.