Inducing Context Gazetteers from Encyclopedic Databases for Named Entity Recognition

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

[1] Suda Lab., Graduate School of Information Science and Technology, the University of Tokyo, Tokyo, Japan
[2] Inui and Okazaki Lab., Graduate School of Information Science, Tohoku University, Sendai, Japan
[3] Japan Science and Technology Agency (JST)
hccho@is.s.u-tokyo.ac.jp
{okazaki,inui}@ecei.tohoku.ac.jp

Abstract. Named entity recognition (NER) is a fundamental task for mining valuable information from unstructured and semi-structured texts. State-of-the-art NER models mostly employ a supervised machine learning approach that heavily depends on local contexts. However, results of recent research have demonstrated that non-local contexts at the sentence or document level can help advance the improvement of recognition performance. As described in this paper, we propose the use of a context gazetteer, the list of contexts with which entity names can co-occur, as new non-local context information. We build a context gazetteer from an encyclopedic database because manually annotated data are often too few to extract rich and sophisticated context patterns. In addition, dependency path is used as sentence level non-local context to capture more syntactically related contexts to entity mentions than linear context in traditional NER. In the discussion of experimentation used for this study, we build a context gazetteer of gene names and apply it for a biomedical NER task. High confidence context patterns appear in various forms. Some are similar to a predicate–argument structure whereas some are in unexpected forms. The experiment results show that the proposed model using both entity and context gazetteers improves both precision and recall over a strong baseline model, and therefore the usefulness of the context gazetteer.

1 Introduction

Named entity recognition (NER) is a task that recognizes the mentions of entities of interest. Entity types vary depending on the target domains. In the general domain, for example, the names of people, locations and organizations are most common entity types [5,25], whereas the names of genes and gene products are in the biomedical domain [12,22]. In fact, NER has been regarded as a fundamental sub-task in many natural language processing (NLP) applications such as information extraction, question and answering, and machine translation.

J. Pei et al. (Eds.): PAKDD 2013, Part I, LNAI 7818, pp. 378–389, 2013.

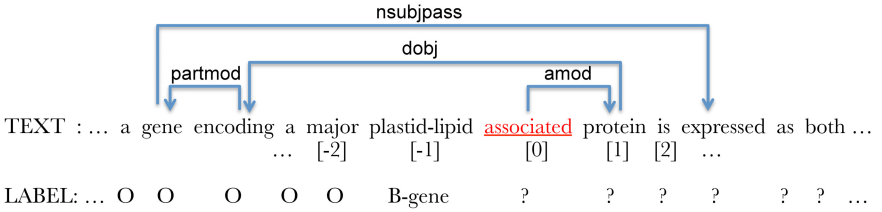

Fig. 1. Example of local and non-local context window. The local context window [-2,2] is shown under the text, whereas the non-local context window is shown with directed arrows. "plastid-lipid associated protein" is the name of a gene where the first word is labeled with the *B-gene* meaning that this is the beginning of a gene name. The dependency label *amod* stands for adjectival modifier, *dobj* for direct object, *partmod* for participial modifier and *nsubjpass* for the passive nominal subject.

NER has been tackled in various ways from rule-based to statistical approaches. However, most state-of-the-art NER models formalize it as a sequence labeling task and employ supervised machine learning approaches such as Conditional Random Fields (CRF) and Support Vector Machines (SVMs). To achieve high-performance, a supervised machine learning approach requires a set of features that are well designed to distinguish mentions of entities from others. Commonly used features are local features obtained from a small and linear window (local context hereinafter). For example, presuming that we shall determine the label of the underlined word "associated" in Fig. 1, then 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 (the relative position of each word is shown under the word). These local features contribute to production of strong baseline models [2,8,20]. However, recent studies [4,13] have demonstrated that incorporating features from non-local context can help further improve the recognition performance. In Fig. 1, for instance, direct and indirect head-words of the word "associated" such as "protein", "encoding", "gene", and "expressed" can be useful non-local features.

As described 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 model. A context gazetteer consists of dependency paths of variable lengths to capture more syntactically meaningful contexts than traditional local contexts. Confidence values are assigned to these contexts to reflect how they are likely to appear with entity names. We build a context gazetteer from a huge amount of highly precise and automatically labeled data using an encyclopedic database because manually annotated data are often too few to extract rich and sophisticated context patterns. Therefore, a context gazetteer is expected to help recognize unknown entity names that do not appear in training data, in addition to out-of-vocabulary (OOV) entity names that are not registered in entity gazetteers. In experiment, we build a context gazetteer of gene names and apply it for a biomedical named entity recognition task. It is particularly interesting that top-ranked entries in the created

context gazetteer have various forms. As expected, there are many predicate–argument structure style contexts using domain specific verbal (and nominal) predicates such as "express", "inhibit" and "promote." Moreover, abbreviation, apposition, and conjunction dependencies are frequently included as a part of high confidence context patterns. These contexts can be interpreted as fragments of domain knowledge that appear in stereotypical syntactic structures in texts. The context gazetteer boosted both the precision from 89.06 to 89.32 and the recall from 82.78 to 83.46. As a consequence, the overall F1-score is improved from 85.81 to 86.29.

The remainder of this paper is organized as follows. In Sec. 2, we explain work related to our research. Section 3 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 [16], and apply it to the BioCreative 2 gene name recognition task [22]. The usefulness of a context gazetteer is demonstrated experimentally. Representative output results are analyzed. We show what kinds of context patterns are mined and how they affect a proposed model using the context gazetteer. Section 5 summarizes the contributions of this work, and explains the future work for generalizing learned contexts.

2 Related Work

This section presents a summary of three types of related studies of sentence level non-local features, gazetteer induction and weakly supervised learning.

Sentence level non-local features usually depend on a deep parsing technique. For example, a previous work [7] used the Stanford dependency parser [17] to exploit features such as the head and governor of the noun phrases in a biomedical NER task. A more recent work [23] evaluated the effect of seven different parsers in feature generation for finding base noun phrases including gene names. However, they extract contexts only from training data, whereas we use a large amount of automatically annotated data. As a result, our approach is likely to provide richer and more sophisticated context patterns than their methods.

Gazetteers are invaluable resources for NER tasks, especially for dealing with unknown words that do not appear in training data. They might have the same semantic categories to target entity classes [9], or related classes that are often more fine-grained sub-classes of the target entity classes [20,26]. Word clusters are also useful resources for NER similar to gazetteers. In a related study [18], the Brown clustering algorithm [3] were applied to NER successfully. A more recent work [11] used the dependency relations between verbs and multiword nouns for clustering multiword expressions. However, to the best of our knowledge, all of the related work that we have surveyed produce entity gazetteers (clusters).

The most similar concept to the contexts in this research can be found in the studies related to weakly supervised learning approach. For instance, a bootstrapping method [21] extracts context patterns from unlabeled data using a small set of seed words (entity mentions in case of NER) for a target class. In turn, it extracts new entity mentions using the extracted context patterns, and

repeats this process. However, the quality of context patterns (and also entity mentions) degrades as iteration goes on because it inevitably suffers from semantic drift. In contrast, our method induces a large number of highly precise contexts without a repetitive process by exploiting an encyclopedic database. This approach have become more realistic lately because of many publicly available resources such as Wikipedia[1] and domain-specific databases.

3 Building a Context Gazetteer

A context gazetteer is a confidence assigned list of dependency paths (hereinafter, contexts) of variable length that can co-occur with target entity names. Figure 2 portrays an exemplary context of length 3. It is a high confidence context

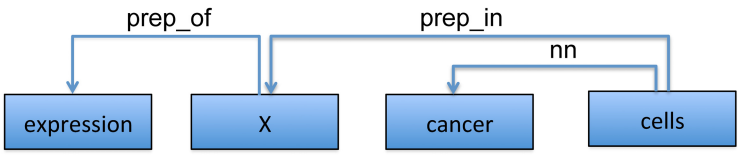

Fig. 2. Example context of the length 3. X is a slot for an entity word. (*pref_of* stands for prepositional modifier of, *pref_in* for prepositional modifier in and *nn* for noun compound modifier.)

in the context gazetteer of gene names that will be used in the experiment section. It means that a word X surrounded by the context consisting of the head word *expression*, a dependent *cells* and a grand-dependent *cancer* with the corresponding dependency relations *prep_of*, *prep_in* and *nn* is likely to be an entity word, which is a part of a target entity name. This context can help to recognize the headword of an underlined gene name in a sentence, "The *expression* of FasL in gastric *cancer cells* and of Fas in apoptotic TIL was also detected in vivo."

A useful context gazetteer should have rich and sophisticated contexts that are specific to target semantic classes. For the first requirement, we extract contexts from a large amount of automatically labeled data rather than a few manually annotated data. To satisfy the second requirement, confidence values are assigned to the extracted contexts. Figure 3 is the flowchart for the context gazetteer generation. Each step is explained in detail in the following.

Step 1. An encyclopedic database consists of domain specific entity names and their descriptions. For each entity name, we label every mention of it in the description using exact string matching. The primary reason for using an encyclopedic database rather than the list of target entity names and some free texts is to

[1] `http://www.wikipedia.org/`

remove the ambiguity of the semantic categories of target entity names appearing in free texts [29]. For example, presuming that we are going to generate labeled data with the names of people using some free text (e.g. newspapers) and a list of the names of people automatically, the process would invariably create very noisy data because human names are often used as the names of companies (e.g., Hewlett-Packard and Ford Motor Company), diseases (e.g. Alzheimer disease), places (e.g., Washington, D.C and St. Paul, Minnesota), and so on.

Step 2. The labeled texts are then parsed. The dependency paths (contexts) involving entity words are extracted. Because of the excessive number of possible contexts, we applied two constraints to context generation. First, the contexts that have no content words (nouns, verbs and adjectives) except an entity word are removed because these contexts are often too general to be effective contexts. Second, we limit the maximum length of contexts depending on the data size.

Step 3. For each context, an entity word is anonymized. Then, contexts can be normalized to increase the coverage of a context gazetteer. For example, stems (or lemmas) are useful instead of words. After normalization, we remove duplicated contexts and keep them unique.

Step 4. Contexts are often ambiguous even if they frequently appear with target entity names. We solve this problem by assigning confidence to each context. Presuming that data D is annotated automatically with the mentions of T different entity types[2], then, the confidence (conditional probability) of an entity type t given a context c is defined as in

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

where e_t is an entity word of the semantic type $t \in T$ in the data D. The estimated confidence is pessimistic, meaning that they are usually lower than they should be because automatically annotated data have high precision but low recall.

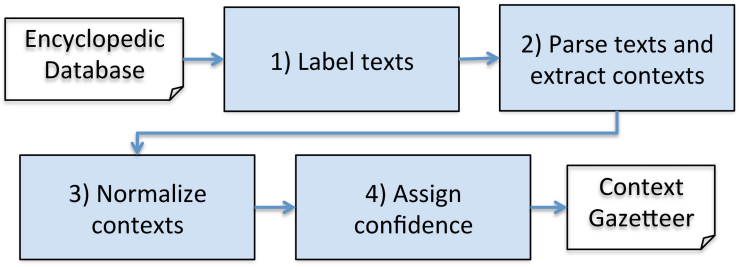

Fig. 3. Building a context gazetteer from an encyclopedic database

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

4 Evaluation

In this section, we create a context gazetteer of gene names from the EntrezGene database [16], and apply it to the BioCreative 2 gene name recognition task [22]. We analyze the effect of the context gazetteer by comparing the NER models with and without the context gazetteer.

4.1 Data

Context Gazetteer. For gazetteer generation, we use the gene names (including synonyms) and the human curated reference information in the EntrezGene. At the first step in Fig. 3, 358,049 abstracts are extracted from the MEDLINE database[3] using reference information. Each abstract is labeled using the gene names referenced in the abstract. The labeled gene names are highly precise because explicit references exist between the gene names and the abstracts.

Second, the labeled texts are parsed using the Stanford POS tagger [27] and dependency parser [17] included in the CoreNLP tool[4]. Then, we extracted the dependency paths (contexts) that involve entity words. Contexts that have no content words aside from entity words are filtered out. The maximum length is set to 5 experimentally.

Third, the entity words of the contexts are anonymized. In the biomedical domain, many entity names include symbols and numbers. For domain-specific normalization, continuous numbers and symbols of the words in the contexts are converted into a representative number (0) and symbol (under-bar), respectively. Lastly, confidence values are assigned to each context using Eq. 1. Contexts appearing less than 10 times are removed in this process because the estimated confidence might be unreliable.

Several extracted contexts having high confidence are presented in Table 1. At the beginning of this study, we expected to obtain contexts similar to predicate-argument structure (PAS) and domain specific relations. For example, the second context in this table indicates that X is likely to be a gene if it appears in a relation with *C-jun* as in "... interaction between X and C-Jun". The fourth and seventh contexts are in the form of PAS using nominal and verbal predicates respectively. 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 ..."

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

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

Table 1. Examples of high confidence extracted context patterns. Conf. stands for confidence. (X is a place-holder, *nsubj* is nominal subject, *conj_and* is conjunction and, *nn* is noun compound modifier, *amod* is adjectival modifier, *dobj* is direct object, and *nsubjpass* is passive nominal subject.)

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)

Entity Gazetteer. We use four entity gazetteers compiled from the EntrezGene, Universal Protein Resource (UniProt) [6], Unified Medical Language System (UMLS) [1] and the Open Biological and Biomedical Ontologies (OBO)[5]. 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 lowercased. This process also applies to the input texts.

For the entity gazetteers compiled from the EntrezGene and the UniProt, we use the single semantic categories: gene and protein. However, the UMLS and the OBO gazetteers have multiple categories, some of which are related to gene names such as peptides and amino acids, but many of which are different biomedical entity categories. During NER system development, we found that not only gene-related categories but also other categories are beneficial for increasing performance.

GENETAG corpus. The BioCreative 2 gene mention recognition task uses the GENETAG corpus [24] comprising 20,000 sentences, of which 15,000 sentences were used for training and 5,000 sentences were used for testing.

We processed raw texts to obtain additional syntactic information for use in feature generation. Raw texts consisting of sentences are split into tokens using a fine-grained tokenization scheme that uses whitespace and non-alphanumeric characters as token boundary markers. When a string is tokenized at non-alphanumeric character, this character also becomes a single character token (e.g., "p53-activated" to "p53", "-" and "activated"). Next, the tokenized text is fed to the GENIA tagger [28] for lemmatization, POS-tagging, and chunking. For each entity gazetteer, the sequences of tokens that appear in the gazetteer are tagged using the BIO labels (e.g., "EntityGaz_B-EntrezGene", "EntityGaz_B-UniProt", etc.). Lastly, for the EntrezGene context gazetteer, the tokens

[5] `http://www.obofoundry.org/`

Table 2. Features used for experiments. Ortho. stands for orthographical features, E. gaz. for entity gazetteer and C. gaz. for context gazetteer.

Class	Description
Token	$\{w_{t-2}, .., w_{t+2}\} \wedge y_t$, $\{w_{t-2,t-1}, .., w_{t+1,t+2}\} \wedge y_t$, $\{\bar{w}_{t-2}, .., \bar{w}_{t+2}\} \wedge y_t$ $\{\bar{w}_{t-2,t-1}, .., \bar{w}_{t+1,t+2}\} \wedge y_t$,
Lemma	$\{l_{t-2}, .., l_{t+2}\} \wedge y_t$, $\{l_{t-2,t-1}, .., l_{t+1,t+2}\} \wedge y_t$, $\{\bar{l}_{t-2}, .., \bar{l}_{t+2}\} \wedge y_t$, $\{\bar{l}_{t-2,t-1}, .., \bar{l}_{t+1,t+2}\} \wedge y_t$
POS	$\{p_{t-2}, .., p_{t+2}\} \wedge y_t$, $\{p_{t-2,t-1}, .., p_{t+1,t+2}\} \wedge y_t$,
Lemma & POS	$\{l_{t-2}p_{t-2}, .., l_{t+2}p_{t+2}\} \wedge y_t$, $\{l_{t-2,t-1}p_{t-2,t-1}, .., l_{t+1,t+2}p_{t+1,t+2}\} \wedge y_t$
Chunk	$\{c_t, w_{t\_last}, \bar{w}_{t\_last}, the_{lhs}\} \wedge y_t$
Character	Character 2,3,4-grams of w_t
Ortho.	All capitalized, all numbers, contain Greek letters, ... (Refer to [15] for the detailed explanation)
E. gaz.	$\{ge_{t-2}, .., ge_{t+2}\} \wedge y_t$, $\{ge_{t-2,t-1}, .., ge_{t+1,t+2}\} \wedge y_t$, $\{ge_{t-2}l_{t-2}, .., ge_{t+2}l_{t+2}\} \wedge y_t$, $\{ge_{t-2,t-1}l_{t-2,t-1}, .., ge_{t+1,t+2}l_{t+1,t+2}\} \wedge y_t$,
C. gaz.	$\{gc_t \wedge y_t\}$

surrounded by the contexts of the gazetteer are 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.

4.2 Machine Learning and Features

For machine learning, we use the CRFsuite [19], which implements first-order linear-chain Conditional Random Fields [14]. 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. The best performing one is chosen.

A set of features used in the experiment is presented in Table 2, and the symbols are explained in Table 3.

4.3 Experiment Results

Table 4 shows an experiment result obtained using various combinations of the four entity gazetteers and the context gazetteer. The numbers in a pair of parentheses show improvement from the model 0 (baseline model) using no gazetteers.

When the context gazetteer is used in combination with the entity gazetteer(s), both precision and recall increase, as shown in models 3 and 5. Considering that precision and recall are tradeoff measures, the experiment result demonstrates the usefulness of the context gazetteer. In addition, the context gazetteer improves recall notably. This is an important merit because NER models usually

Table 3. Symbols used for features (see Table 2)

Symbol	Description
w_t	A *t-th* word
$\bar{w}_t$	A normalized *t-th* word where successive numbers and symbols are converted into a single zero and under-bar
l_t	A *t-th* lemma
$\bar{l}_t$	A normalized *t-th* lemma
p_t	A *t-th* POS-tag
c_t	The chunk type of w_t
w_{t_last}	The last word of a current chunk
$\bar{w}_{t\_last}$	The normalized last word of a current chunk
the_{lhs}	True if 'the' exists from the beginning of a current chunk to w_{t-1}
ge_t	Entity gazetteer label for the *t-th* word
gc_t	Context gazetteer label for the *t-th* word

Table 4. Performance evaluation using entity and context gazetteers. ALL means the gazetteers compiled from the EntrezGene, UniProt, UMLS, and OBO databases.

Model #	Entity Gaz.	Context Gaz.	Precision	Recall	F1-score
0	None	None	87.99 (+0.00)	81.71 (+0.00)	84.73 (+0.00)
1	None	EntrezGene	88.06 (+0.07)	81.42 (-0.29)	84.61 (-0.12)
2	EntrezGene	None	88.54 (+0.55)	82.17 (+0.46)	85.24 (+0.51)
3	EntrezGene	EntrezGene	88.66 (+0.67)	82.99 (+1.28)	85.73 (+1.00)
4	ALL	None	89.06 (+1.07)	82.78 (+1.07)	85.81 (+1.08)
5	ALL	EntrezGene	89.32 (+1.33)	83.46 (+1.75)	86.29 (+1.56)

exhibit high precision but low recall [10] because of the asymmetric data where one class label, O, dominates all other classes.

Surprisingly, when only the context gazetteer is used, the overall performance drops slightly. We suspect that some relation exists between entity gazetteers and context gazetteers but further investigation is necessary to reveal it.

4.4 Result Analysis

We manually compared about 20% of the output of models 4 and 5 to see how the context gazetteer features affect the tagging results.

There are 32 gene names correctly recognized by model 5 but not by model 4. In all of these cases, one or more context gazetteer features are triggered. The following list shows several examples in which model 5 recognized the under-barred gene names and model 4 recognized the italicized gene names.

- One major transcript encodes MEQ, a *339-amino-acid bZIP protein* which is homologous to the Jun/Fos family of transcription factors.
- The association of I-92 with *p92*, *p84*, *p75*, *p73*, *p69*, and *p57* was completely reversible after treatment with the detergent deoxycholate (DOC).

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

Two context gazetteer features are triggered for the gene name "MEQ", "dobj(encode, X)", and "appos(X, protein)." 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 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 is a normalized number. Contexts of these kinds are the fragments of domain specific knowledge and usually have high confidence (0.5 for this context). In the last example, the gene name "IP-30" has a context gazetteer feature "prep_of(function, X)" and a more specific one "nsubjpass(known, function), prep_of(function, X)" with confidence 0.44 and 0.54. These contexts can be interpreted as domain-specific expressions where figuring out the function of a gene is a much more important task than others (54% vs. the rest).

However, 15 gene names are recognized by model 4, but not by model 5. Context gazetteer features are not triggered for 3 cases. Because we use the words (not stems or lemmas) in the contexts, the coverage might be not sufficiently high. For the other 12 cases, context gazetteer features are fired, but these gene names are not recognized. We are currently investigating the causes of these cases.

5 Conclusion and Future Work

As described in this paper, we proposed the use of a context gazetteer as a new non-local feature for NER. We also described how to induce a rich and sophisticated context gazetteer from automatically annotated data using an encyclopedic database. Compared to the feature aggregation methods [4,13,20], 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. The proposed method is applied to a biomedical NER task. Its usefulness is demonstrated in addition to entity gazetteers.

However, we also uncovered difficulties. First, 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. Second, gene names (or parts of them) often appear as a part of contexts. Although these contexts often have very high confidence, they may not be general patterns. They can be more useful if they were replaced by some general gene name wildcards.

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

References

1. Bodenreider, O.: The unified medical language system (umls): integrating biomedical terminology. Nucleic Acids Research 32(suppl. 1), D267–D270 (2004)
2. Borthwick, A., Sterling, J., Agichtein, E., Grishman, R.: Nyu: Description of the mene named entity system as used in muc-7. In: Proceedings of the Seventh Message Understanding Conference, MUC-7 (1998)
3. Brown, P.F., de Souza, P.V., Mercer, R.L., Pietra, V.J.D., Lai, J.C.: Class-based n-gram models of natural language. Journal of Computational Linguistics 18(4), 467–479 (1992)
4. Chieu, H.L., Ng, H.T.: Named entity recognition with a maximum entropy approach. In: Proceedings of the Seventh CoNLL at HLT-NAACL 2003, vol. 4, pp. 160–163 (2003)
5. Chinchor, N.A.: Overview of MUC-7/MET-2. In: Proceedings of the Seventh Message Understanding Conference (MUC7) (April 1998)
6. Consortium, T.U.: Reorganizing the protein space at the universal protein resource (uniprot). Nucleic Acids Research 40(D1), D71–D75 (2012)
7. Finkel, J., Dingare, S., Nguyen, H., Nissim, M., Manning, C., Sinclair, G.: Exploiting context for biomedical entity recognition: from syntax to the web. In: Proceedings of the International Joint Workshop on NLPBA, pp. 88–91 (2004)
8. Finkel, J.R., Grenager, T., Manning, C.: Incorporating non-local information into information extraction systems by gibbs sampling. In: Proceedings of the 43rd Annual Meeting on ACL, pp. 363–370 (2005)
9. Florian, R., Ittycheriah, A., Jing, H., Zhang, T.: Named entity recognition through classifier combination. In: Proceedings of the Seventh CoNLL at HLT-NAACL 2003, vol. 4, pp. 168–171 (2003)
10. Kambhatla, N.: Minority vote: at-least-n voting improves recall for extracting relations. In: Proceedings of COLING-ACL, pp. 460–466 (2006)
11. Kazama, J., Torisawa, K.: Inducing Gazetteers for Named Entity Recognition by Large-Scale Clustering of Dependency Relations. In: Proceedings of ACL-HLT, pp. 407–415 (2008)
12. Kim, J.D., Pyysalo, S., Ohta, T., Bossy, R., Nguyen, N., Tsujii, J.: Overview of bionlp shared task 2011. In: Proceedings of the BioNLP Shared Task 2011 Workshop, pp. 1–6 (2011)
13. Krishnan, V., Manning, C.D.: An effective two-stage model for exploiting non-local dependencies in named entity recognition. In: Proceedings of COLING-ACL, pp. 1121–1128 (2006)
14. Lafferty, J.D., McCallum, A., Pereira, F.C.N.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proceedings of the Eighteenth International Conference on Machine Learning, pp. 282–289 (2001)
15. Lee, K.J., Hwang, Y.S., Kim, S., Rim, H.C.: Biomedical named entity recognition using two-phase model based on svms. Journal of Biomedical Informatics 37(6), 436–447 (2004)
16. Maglott, D., Ostell, J., Pruitt, K.D., Tatusova, T.: Entrez gene: Gene-centered information at ncbi. Nucleic Acids Research 33(suppl. 1), D54–D58 (2005)
17. Marneffe, M.C.D., MacCartney, B., Manning, C.D.: Generating typed dependency parses from phrase structure parses. In: Proceedings of LREC 2006 (2006)
18. Miller, S., Guinness, J., Zamanian, A.: Name tagging with word clusters and discriminative training. In: Susan Dumais, D.M., Roukos, S. (eds.) Proceedings of HLT-NAACL, May 2-May 7, pp. 337–342 (2004)

19. Okazaki, N.: Crfsuite: A fast implementation of conditional random fields, crfs (2007), http://www.chokkan.org/software/crfsuite/
20. Ratinov, L., Roth, D.: Design challenges and misconceptions in named entity recognition. In: Proceedings of the Thirteenth Conference on CoNLL, pp. 147–155 (2009)
21. Riloff, E., Shepherd, J.: A corpus-based approach for building semantic lexicons. In: Proceedings of the Second Conference on EMNLP, pp. 117–124 (1997)
22. Smith, L., Tanabe, L., Ando, R., Kuo, C.J., Chung, I.F., Hsu, C.N., Lin, Y.S., Klinger, R., Friedrich, C., Ganchev, K., Torii, M., Liu, H., Haddow, B., Struble, C., Povinelli, R., Vlachos, A., Baumgartner, W., Hunter, L., Carpenter, B., Tsai, R., Dai, H.J., Liu, F., Chen, Y., Sun, C., Katrenko, S., Adriaans, P., Blaschke, C., Torres, R., Neves, M., Nakov, P., Divoli, A., Mana-Lopez, M., Mata, J., Wilbur, W.J.: Overview of biocreative ii gene mention recognition. Genome Biology 9(suppl. 2), S2 (2008)
23. Smith, L.H., Wilbur, W.J.: Value of parsing as feature generation for gene mention recognition. Journal of Biomedical Informatics 42(5), 895–904 (2009)
24. Tanabe, L., Xie, N., Thom, L., Matten, W., Wilbur, W.J.: Genetag: a tagged corpus for gene/protein named entity recognition. BMC Bioinformatics 6(suppl. 1), S3 (2005)
25. Tjong Kim Sang, E.F., De Meulder, F.: Introduction to the conll-2003 shared task: language-independent named entity recognition. In: Proceedings of the Seventh CoNLL at HLT-NAACL 2003, vol. 4, pp. 142–147 (2003)
26. Torisawa, K.: Exploiting wikipedia as external knowledge for named entity recognition. In: Proceedings of the Joint Conference on EMNLP-CoNLL, pp. 798–707 (2007)
27. Toutanova, K., Klein, D., Manning, C.D., Singer, Y.: Feature-rich part-of-speech tagging with a cyclic dependency network. In: Proceedings of the 2003 Conference of the HLT-NAACL, vol. 1, pp. 173–180 (2003)
28. Tsuruoka, Y., Tsujii, J.: Bidirectional inference with the easiest-first strategy for tagging sequence data. In: Proceedings of the Conference on HLT-EMNLP, pp. 467–474 (2005)
29. Usami, Y., Cho, H.C., Okazaki, N., Tsujii, J.: Automatic acquisition of huge training data for bio-medical named entity recognition. In: Proceedings of BioNLP 2011 Workshop, pp. 65–73 (2011)