Building a Corpus for Japanese Wikification with Fine-Grained Entity Classes

Davaajav Jargalsaikhan Naoaki Okazaki Koji Matsuda Kentaro Inui
Tohoku University, Japan
{davaajav, okazaki, matsuda, inui}@ecei.toholu.ac.jp

Abstract

In this research, we build a Wikification corpus for advancing Japanese Entity Linking. This corpus consists of 340 Japanese newspaper articles with 25,675 entity mentions. All entity mentions are labeled by a fine-grained semantic classes (200 classes), and 19,121 mentions were successfully linked to Japanese Wikipedia articles. Even with the fine-grained semantic classes, we found it hard to define the target of entity linking annotations and to utilize the fine-grained semantic classes to improve the accuracy of entity linking.

1 Introduction

Entity linking (EL) recognizes mentions in a text and associates them to their corresponding entries in a knowledge base (KB), for example, Wikipedia[1], Freebase (Bollacker et al., 2008), and DBPedia (Lehmann et al., 2015). In particular, when linked to Wikipedia articles, the task is called Wikifiation (Mihalcea and Csomai, 2007). Let us consider the following sentence.

> On the 2nd of June, the team of Japan will play World Cup (W Cup) qualification match against Honduras in the second round of Kirin Cup at Kobe Wing Stadium, the venue for the World Cup.

Wikification is expected to link "Soccer" to the Wikipedia article titled *Soccer*, "World Cup" and "W Cup"[2] to *FIFA World Cup 2002*, "team of Japan" to *Japan Football Team*, "Kobe City" to *Kobe*, "Kobe Wing Stadium" to *Misaki Park Stadium*. Since there is no entry for "Second Round of Kirin Cup", the mention is labeled as NIL.

[1] https://www.wikipedia.org/

[2] "W Cup" is a Japanese-style abbreviation of "World Cup".

EL is useful for various NLP tasks, e.g., Question-Answering (Khalid et al., 2008), Information Retrieval (Blanco et al., 2015), Knowledge Base Population (Dredze et al., 2010), Co-Reference Resolution (Hajishirzi et al., 2013). There are about a dozen of datasets targeting EL in English, including UIUC datasets (ACE, MSNBC) (Ratinov et al., 2011), AIDA datasets (Hoffart et al., 2011), and TAC-KBP datasets (2009–2012 datasets) (McNamee and Dang, 2009).

Ling et al. (2015) discussed various challenges in EL. They argued that the existing datasets are inconsistent with each other. For instance, TAC-KBP targets only mentions belonging to PERSON, LOCATION, ORGANIZATION classes. Although these entity classes may be dominant in articles, other tasks may require information on natural phenomena, product names, and institution names. In contrast, the MSNBC corpus does not limit entity classes, linking mentions to any Wikipedia article. However, the MSNBC corpus does not have a NIL label even if a mention belongs to an important class such as PERSON or LOCATION, unlike the TAC-KBP corpus.

There are few studies addressing on Japanese EL. Furukawa et al. (2014) conducted a study on recognizing technical terms appearing in academic articles and linking them to English Wikipedia articles. Hayashi et al. (2014) proposed an EL method that simultaneously performs both English and Japanese Wikification, given parallel texts in both languages. Nakamura et al. (2015) links keywords in social media into English Wikipedia, aiming at a cross-language system that recognizes topics of social media written in any language. Osada et al. (2015) proposed a method to link mentions in news articles for organizing local news of different prefectures in Japan.

However, these studies do not necessarily ad-

Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics – Student Research Workshop, pages 138–144,
Berlin, Germany, August 7-12, 2016.

vance EL on a Japanese KB. As of January 2016, Japanese Wikipedia and English Wikipedia include about 1 million and 5 million, respectively, articles. However, there are only around 0.56 million inter-language links between Japanese and English. Since most of the existing KBs (e.g., Freebase and DBPedia) originate from Wikipedia, we cannot expect that English KBs cover entities that are specific to Japanese culture, locals, and economics. Moreover, a Japanese EL system is useful for populating English knowledge base as well, harvesting source documents written in Japanese.

To make matters worse, we do not have a corpus for Japanese EL, i.e., Japanese mentions associated with Japanese KB. Although (Murawaki and Mori, 2016) concern with Japanese EL, the corpus they have built is not necessarily a corpus for Japanese EL. The motivation behind their work comes from the difficulty of word segmentation for unsegmented languages, like Chinese or Japanese. (Murawaki and Mori, 2016) approach the word segmentation problem from point of view of Wikification. Their focus is on the word segmentation rather than on the linking.

In this research, we build a Japanese Wikification corpus in which mentions in Japanese documents are associated with Japanese Wikipedia articles. The corpus consists of 340 newspaper articles from Balanced Corpus of Contemporary Written Japanese (BCCWJ) [3] annotated with fine-grained named entity labels defined by Sekine's Extended Named Entity Hierarchy (Sekine et al., 2002) [4].

2 Dataset Construction

To give a better understanding of our dataset we briefly compare it with existing English datasets. The most comparable ones are UIUC (Ratinov et al., 2011) and TAC-KBP 2009–2012 datasets (McNamee and Dang, 2009). Although, AIDA datasets are widely used for Disambiguation of Entities, AIDA uses YAGO, an unique Knowledge Base derived from Wikipedia, GeoNames and Wordnet, which makes it difficult to compare. UIUC is similar to our dataset in a sense that it links to any Wikipedia article without any semantic class restrictions, unlike TAC-KBP which is limited to mentions that belong to PERSON, LOCATION or ORGANIZATION classes only. When an article is not present in Wikipedia, UIUC does not record this information in any way. On the contrary, TAC-KBP[5] and our datasets have NIL tag used to mark a mention when it does not have an entry in KB.

2.1 Design Policy

Ling et al. (2015) argued that the task definition of EL itself is challenging: whether to target only named entities (NEs) or to include general nouns; whether to limit semantic classes of target NEs; how to define NE boundaries; how specific the links should be; and how to handle metonymy.

The original (Hashimoto et al., 2008) corpus is also faced with similar challenges: mention abbreviations that result in the string representation that is an exact match to the string representation of another mention, abbreviated or not (for example, "Tokyo (City)" and "TV Tokyo"), metonymy and synecdoche.

As for the mention "World Cup" in the example in Section 1, we have three possible candidates entities, *World Cup*, *FIFA World Cup*, and *2002 FIFA World Cup*. Although all of them look reasonable, *2002 FIFA World Cup* is the most suitable, being more specific than others. At the same time, we cannot expect that Wikipedia includes the most specific entities. For example, let us suppose that we have a text discussing a possible venue for 2034 FIFA World Cup. As of January 2016, Wikipedia does not include an article about 2034 FIFA World Cup[6]. Thus, it may be a difficult decision whether to link it to *FIFA World Cup* or make it NIL.

Moreover, the mention "Kobe Wing Stadium" includes nested NE mentions, "Kobe (City)" and "Kobe Wing Stadium". Furthermore, although the article titled "Kobe Wing Stadium" does exist in Japanese Wikipedia, the article does not explain the stadium itself but explains the company running the stadium. Japanese Wikipedia includes a separate article *Misaki Park Stadium* describing the stadium. In addition, the mention "Honduras" does not refer to Honduras as a country, but as the national soccer team of Honduras.

In order to separate these issues raised by NEs

[3] http://pj.ninjal.ac.jp/corpus_center/bccwj/en/

[4] https://sites.google.com/site/extendednamedentityhierarchy/

[5] TAC-KBP 2012 requires NIL to be clustered in accordance to the semantic classes.

[6] Surprisingly, Wikipedia includes articles for the future World Cups up to 2030.

from the EL task, we decided to build a Wikification corpus on top of a portion of BCCWJ corpora with Extended Named Entity labels annotated (Hashimoto et al., 2008). This corpus consists of 340 newspaper articles where NE boundaries and semantic classes are annotated. This design strategy has some advantages. First, we can omit the discussion on semantic classes and boundaries of NEs. Second, we can analyze the impact of semantic classes of NEs to the task of EL.

2.2 Annotation Procedure

We have used *brat rapid annotation tool* (Stenetorp et al., 2012) to effectively link mentions to Wikipedia articles. Brat has a functionality of importing external KBs (e.g., Freebase or Wikipedia) for EL. We have prepared a KB for Brat using a snapshot of Japanese Wikipedia accessed on November 2015. We associate a mention to a Wikipedia ID so that we can uniquely locate an article even when the title of the article is changed. We configure Brat so that it can present a title and a lead sentence (short description) of each article during annotation.

Because this is the first attempt to build a Japanese Wikification dataset on a fine-grained NE corpus, we did not limit the semantic classes of target NEs in order to analyze the importance of different semantic classes. However, based on preliminary investigation results, we decided to exclude the following semantic classes from targets of the annotation: `Timex` (Temporal Expression, 12 classes), `Numex` (Numerical Expression, 34 classes), `Address` (e.g., postal address and urls, 1 class), `Title_Other` (e.g., *Mr.*, *Mrs.*, 1 class), `Facility_Part` (e.g, *9th floor*, *second basement*, 1 class). Mentions belonging to other classes were linked to their corresponding Wikipedia pages.

We asked three Japanese native speakers to link mentions into Wikipedia articles using Brat. We gave the following instructions to obtain consistent annotations:

1. Choose the entity that is the most specific in possible candidates.
2. Do not link a mention into a disambiguation page, category page, nor WikiMedia page.
3. Link a mention into a section of an article only when no suitable article exists for the mention.

Attribute	Value
# articles	340
# mentions	25,675
# links	19,121
# NILs	6,554
# distinct mentions	7,118
# distinct entities	6,008

Table 1: Statistics of the corpus built by this work.

Annotator pair	Agreement
Annotators 1 and 2	0.910
Annotators 2 and 3	0.924

Table 2: Inter-annotator agreement.

2.3 Annotation Results

Table 1 reports the statistics of the corpus built by this work. Out of 25,675 mentions satisfying the conditions explained in Section 2.2, 19,121 mentions were linked to Japanese Wikipedia articles. In total, 7,118 distinct mentions were linked to 6,008 distinct entities. Table 2 shows the high inter- annotator agreement (the Cohen-Kappa's coefficient) of the corpus[7].

In order to find important/unimportant semantic classes of NEs for EL, we computed the link rate for each semantic class. Link rate of a semantic class is the ratio of the number of linkable (non-NIL) mentions belonging that class to the total number of mentions of that class occurring throughout the corpus. Table 3 presents semantic classes with the highest and lowest link rates[8]. Popular NE classes such as `Province` and `Pro_Sports_Organization` had high link rates. Semantic classes such as Book and `Occasion_Other` had low link rates because these entities are rare and uncommon. However, we also found it difficult to limit the target of entity linking based only on semantic classes because the importance of the semantic classes with

[7]We cannot compute the inter-annotator agreement between Annotators 1 and 3, who have no overlap articles for annotation.

[8]In this analysis, we removed semantic classes appearing less than 100 times in the corpus. We concluded that those minor semantic classes do little help in revealing the nature of the dataset we have built. Most of them had perfect or near to zperfect link rates with mentions being rare and uniquely identifiable.

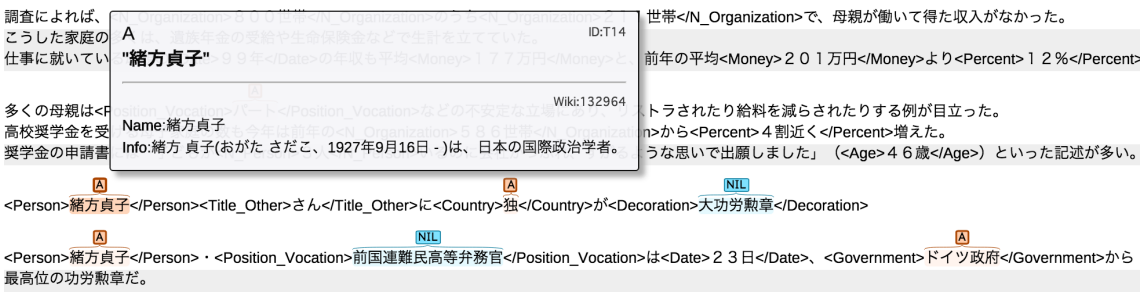

Figure 1: A screenshot of the annotation senvironment with Brat.

low link rates depends on the application; for example, `Occasion_Other`, which has the lowest link rate, may be crucial for event extraction from text.

In the research on word sense disambiguation (WSD), it is common to assume that the identical expressions have the same sense throughout the text. This assumption is called *one-sense-per-discourse*. In our corpus, 322 out of 340 (94.7%) articles satisfy the assumption. A few instances include: expressions "Bush" referred to both *George H. W. Bush* and *George W. Bush* (the former is often referred as Bush Senior and the latter as Bush Junior); and expressions "Tokyo" referred to both *Tokyo Television* and *Tokyo city*.

2.4 Difficult Annotation Cases

We report cases where annotators found difficult to choose an entity from multiple potential candidates. Mention boundaries from the original corpus are indicated by underline.

Nested entities

> It was assumed that the role initially served as a temporary peacemaker to persuade Ali al-Sistani, the spiritual leader of Shia Muslims: `Position_Vocation`.

Since the mention in the sentence refers to the highest ranking position of a specific religion, it is inappropriate to link the mention to the article *Spiritual Leader* nor *Shia Muslim*. Therefore, we decided to mark this mention as NIL.

Entity changes over time

> In his greeting speech, the representative Ito expressed his opinion on the upcoming gubernatorial election: `Event_Other` and Sapporo city mayoral election.

This article was about the Hokkaido Prefecture gubernatorial election held in 2003. Since the BCCWJ corpus does not provide timestamps of articles, it is difficult to identify the exact event. However, this article has a clue in another place, "the progress of the developmental project from 2001". For this reason, the annotators could resolve the mention to *2003 Hokkaido Prefecture gubernatorial election*. Generally, it is difficult to identify events that are held periodically. The similar issue occurs in mentions regarding position/profession (e.g., "former president") and sport events (e.g., "World Cup").

Japanese EL is similar to English El: the same challenges of mention ambiguity (nested entities, metonymy) still persist. With the Japanese Wikification, a variation of the task that takes advantage of the cross-lingual nature of Wikipedia is worth exploring.

3 Wikification Experiment

In this section, we conduct an experiment of Wikification on the corpus built by this work. Wikification is decomposed into two steps: recognizing a mention m in the text, and predicting the corresponding entity e for the mention m. Because the corpus was built on the corpus with NE mentions recognized, we omit the step of entity mention recognition.

3.1 Wikification without fine-grained semantic classes

Our experiment is based on the disambiguation method that uses the probability distribution of anchor texts (Spitkovsky and Chang, 2012). Given a mention m, the method predicts an entity $\hat{e}$ that yields the highest probability $p(e|m)$,

$$\hat{e} = \underset{e \in E}{\operatorname{argmax}}\, p(e|m). \quad (1)$$

Category	Example	Link Rate	# of Links	# of Occurrences
Province	Fukuoka Prefecture	0.983	678	690
Country	United States of America	0.976	1924	1964
GPE_Other	Ginza	0.974	115	118
Political_Party	Liberal Democratic Party	0.967	236	244
Pro_Sports_Organization	Yomiuri Giants	0.997	290	300
City	Sendai City	0.947	1354	1430
Company_Group	JR	0.928	103	111
Mammal	Kangaroo	0.906	164	181
International_Organization	NATO	0.891	188	211
Company	NTT	0.883	647	733
	...			
Game	Summer Olympics	0.576	167	290
Conference	34th G8 summit	0.548	74	135
Public_Institution	Takatsuki City Office	0.451	105	233
Book	Sazae-san	0.412	49	119
Political_Organization_Other	Takeshita faction	0.407	68	167
Organization_Other	General Coordination Division	0.393	55	140
GOE_Other	White House	0.363	99	274
Plan	Income Doubling Plan	0.273	32	117
Character	Mickey Mouse	0.145	29	200
Occasion_Other	Tsukuba EXPO	0.113	28	226

Table 3: 10 classes with the highest and the lowest link rates among the classes that occurred more than 100 times

Here, E is the set of all articles in Japanese Wikipedia. The conditional probability $p(e|m)$ is estimated by the anchor texts in Japanese Wikipedia,

$$p(e|m) = \frac{\text{\# occurrences of } m \text{ as anchors to } e}{\text{\#occurrences of } m \text{ as anchors}}. \quad (2)$$

If $\forall e : p(e|m) = 0$ for the mention m, we mark the mention as NIL. Ignoring contexts of mentions, this method relies on the popularity of entities in the anchor texts of the mention m. The accuracy of this method was 53.31% (13,493 mentions out of 25,309).

3.2 Wikification with fine-grained semantic classes

Furthermore, we explore the usefulness of the fine-grained semantic classes for Wikification. This method estimates probability distributions conditioned on a mention m and its semantic class c. Idealy, we would like to predict an entity $\hat{e}$ with,

$$\hat{e} = \underset{e \in E, c \in C}{\operatorname{argmax}} p(e|m, c) \quad (3)$$

However, it is hard to estimate the probability distribution $p(e|m, c)$ directly from the Wikipedia articles. Instead, we decompose $p(e|m, c)$ into $p(e|m)p(e|c)$ to obtain,

$$\hat{e} = \underset{e \in E, c \in C}{\operatorname{argmax}} p(e|m)p(e|c). \quad (4)$$

Here, C is the set of all semantic classes included in Sekine's Extended Named Entity Hierarchy. In addition, we apply Bayes' rule to $p(e|c)$,

$$\hat{e} = \underset{e \in E, c \in C}{\operatorname{argmax}} p(e|m)p(c|e)p(e) \quad (5)$$

The probability distribution $p(c|e)$ bridges Wikipedia articles and semantic classes defined in Sekine's Extended Named Entity Hierarchy. We adapt a method to predict a semantic class of a Wikipedia article (Suzuki et al., 2016) for estimating $p(c|e)$. The accuracy of this method was 53.26% (13,480 mentions out of 25,309), which is slightly lower than that of the previous method. The new method improved 627 instances mainly with LOCATION Category (e.g., country names and city names). For example,

> The venue is Aichi Welfare Pension Hall in Ikeshita, Nagoya
> Semantic Class: City Correct:*Nagoya City*
> Old Method: *Nagoya Station*
> New Method: *Nagoya City*

Because *Nagoya Station* is more popular in anchor texts in Japanese Wikipedia, the old method predicts *Nagoya Station* as the entity for the mention *Nagoya*. In contrast, the new method could leverage the semantic class, City to avoid the mistake. We could observe similar improvements for distinguishing Country – Language, Person – Location, Location – Sports Team.

However, the new method degraded 664 instances mainly because the fine-grained entity classes tried to map them into too specific entities. More than half of such instances belonged to POSITION_VOCATION semantic class. For example, mention "Prime Minister" was mistakingly mapped to *Prime Minister of Japan* instead of *Prime Minister*.

4 Future Work

In our future work, we will incorporate the context information of the text in the Wikification process and further investigate the definition of the target of entity linking annotations. Although incorporating semantic classes of entities has a potential to improve Wikification quality, some problems still remain even with the semantic classes. Here, we explain some interesting cases.

Name variations

> During the summer, a JASRAC Correct: *Japanese Society for Rights of Authors, Composers and Publishers* Predicted: NIL staff came to the shop to explain it.

This type of mistakes are caused by the lack of aliases and redirects in Wikipedia. In this example, the mention 'JASRAC' was predicted as NIL because Wikipedia did not include JASRAC as an alias for Japanese Society for Rights of Authors, Composers and Publishers.

Link bias in Wikipedia

> Thousands have participated in the funeral held at World Trade Center Correct: *World Trade Center (1973-2001)* Predicted: *World Trade Center (Tokyo)*, which is known as "Ground Zero".

In this example, the mention "World Trade Center" refers to *World Trade Center (1973–2001)* with strong clues in the surrounding context "Ground Zero". Both of the presented methods predict it as *World Trade Center (Tokyo)* because there is a building with the identical name in Japan. Using Japanese Wikipedia articles for estimating the probability distribution, Japanese entities are more likely to be predicted.

5 Conclusion

In this research, we have build a Wikification corpus for advancing Japanese Entity Linking. We have conducted Wikification experiment using using fine grained semantic classes. Although we expect an effect of the fine-grained semantic classes, we could no observe an improvement in terms of the accuracy on the corpus. The definition of the target of entity linking annotations requires further investigation. We are distributing the corpus on the Web site `http://www.cl.ecei.tohoku.ac.jp/jawikify`.

Acknowledgments

This work was partially supported by *Research and Development on Real World Big Data Integration and Analysis*, MEXT and JSPS KAKENHI Grant number 15H05318.

References

Roi Blanco, Giuseppe Ottaviano, and Edgar Meij. 2015. Fast and space-efficient entity linking for queries. In *Proc. of WSDM*, pages 179–188.

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: A collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*, pages 1247–1250.

Mark Dredze, Paul McNamee, Delip Rao, Adam Gerber, and Tim Finin. 2010. Entity disambiguation for knowledge base population. In *Proc. of COLING*, pages 277–285.

Tatsuya Furukawa, Takeshi Sagara, and Akiko Aizawa. 2014. Semantic disambiguation for cross-lingual entity linking (in Japanese). *Journal of Japan Society ofInformation and Knowledge*, 24(2):172–177.

Hannaneh Hajishirzi, Leila Zilles, Daniel S. Weld, and Luke Zettlemoyer. 2013. Joint coreference resolution and named-entity linking with multi-pass sieves. In *Proc. of EMNLP*, pages 289–299.

Taichi Hashimoto, Takashi Inui, and Koji Murakami. 2008. Constructing extended named entity annotated corpora (in Japanese). In *IPSJ Natural Lnaguage Processing (2008-NL-188)*, pages 113–120.

Yoshihiko Hayashi, Kenji Hayashi, Masaaki Nagata, and Takaaki Tanaka. 2014. Improving Wikification of bitexts by completing cross-lingual information. In *The 28th Annual Conference of the Japanese Society for Artificial Intelligence*, pages 1A2–2.

Johannes Hoffart, Mohamed Amir Yosef, Ilaria Bordino, Hagen Fürstenau, Manfred Pinkal, Marc Spaniol, Bilyana Taneva, Stefan Thater, and Gerhard Weikum. 2011. Robust disambiguation of named entities in text. In *Proc. of EMNLP*, pages 782–792.

Mahboob Alam Khalid, Valentin Jijkoun, and Maarten De Rijke. 2008. The impact of named entity normalization on information retrieval for question answering. In *Proc. of ECIR*, pages 705–710.

Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N. Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick van Kleef, Sören Auer, and Christian Bizer. 2015. DBpedia — a large-scale, multilingual knowledge base extracted from Wikipedia. *Semantic Web*, 6(2):167–195.

Xiao Ling, Sameer Singh, and Daniel Weld. 2015. Design challenges for entity linking. *TACL*, 3:315–328.

Paul McNamee and Hoa Trang Dang. 2009. Overview of the TAC 2009 knowledge base population track. In *Text Analysis Conference (TAC)*, pages 111–113.

Rada Mihalcea and Andras Csomai. 2007. Wikify!: Linking documents to encyclopedic knowledge. In *Proc. of CIKM*, pages 233–242.

Yugo Murawaki and Shinsuke Mori. 2016. Wicification for scriptio continua. In *Proc. of LREC*, pages 1346–1351.

Tatsuya Nakamura, Masumi Shirakawa, Takahiro Hara, and Shojiro Nishio. 2015. An entity linking method for closs-lingual topic extraction from social media (in Japanese). In *DEIM Forum 2015*, pages A3–1.

Seiya Osada, Keigo Suenaga, Yoshizumi Shogo, Kazumasa Shoji, Tsuneharu Yoshida, and Yasuaki Hashimoto. 2015. Assigning geographical point information for document via entity linking (in Japanese). In *Proceedings of the Twenty-first Annual Meeting of the Association for Natural Language Processing*, pages A4–4.

Lev Ratinov, Dan Roth, Doug Downey, and Mike Anderson. 2011. Local and global algorithms for disambiguation to Wikipedia. In *Proc. of ACL-HLT*, pages 1375–1384.

Satoshi Sekine, Kiyoshi Sudo, and Chikashi Nobata. 2002. Extended named entity hierarchy. In *Proceedings of LREC 2002*.

Valentin I. Spitkovsky and Angel X. Chang. 2012. A cross-lingual dictionary for english Wikipedia concepts. In *Proc. of LREC*, pages 3168–3175.

Pontus Stenetorp, Sampo Pyysalo, Goran Topic, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. brat: a web-based tool for nlp-assisted text annotation. In *Proceedings of the Demonstrations Session at EACL*.

Masatoshi Suzuki, Koji Matsuda, Satoshi Sekine, Naoaki Okazaki, and Kentaro Inui. 2016. Multi-label classification of wikipedia articles into fine-grained named entity types (in japanese). Proceedings of the Twenty-second Annual Meeting of the Association for Natural Language Processing.