Wikification for Scriptio Continua

Yugo Murawaki, Shinsuke Mori
{Graduate School of Informatics, Academic Center for Computing and Media Studies}, Kyoto University
Yoshida-honmachi, Sakyo-ku, Kyoto, 606-8501, Japan
{murawaki, forest}@i.kyoto-u.ac.jp

Abstract

The fact that Japanese employs *scriptio continua*, or a writing system without spaces, complicates the first step of an NLP pipeline. Word segmentation is widely used in Japanese language processing, and lexical knowledge is crucial for reliable identification of words in text. Although external lexical resources like Wikipedia are potentially useful, segmentation mismatch prevents them from being straightforwardly incorporated into the word segmentation task. If we intentionally violate segmentation standards with the direct incorporation, quantitative evaluation will be no longer feasible. To address this problem, we propose to define a separate task that directly links given texts to an external resource, that is, wikification in the case of Wikipedia. By doing so, we can circumvent segmentation mismatch that may not necessarily be important for downstream applications. As the first step to realize the idea, we design the task of Japanese wikification and construct wikification corpora. We annotated subsets of the Balanced Corpus of Contemporary Written Japanese plus Twitter short messages. We also implement a simple wikifier and investigate its performance on these corpora.

Keywords: corpus annotation, wikification, word segmentation

1. Introduction

Wikification is the task of detecting concept mentions in text and disambiguating them into their corresponding Wikipedia articles. It was originally introduced as an information retrieval task (Mihalcea and Csomai, 2007; Ferragina and Scaiella, 2010), and then has found a wide range of NLP applications including measuring semantic relatedness between text fragments (Gabrilovich and Markovitch, 2007), text classification (Chang et al., 2008), and coreference resolution (Ratinov and Roth, 2012).

In this paper, we present new Japanese corpora for wikification. A major difference from previous studies on English lies in the fact that like Chinese and Thai, Japanese employs *scriptio continua*, or a writing system in which words are not delimited by white space. For this reason, word segmentation is widely used as a prerequisite for most NLP tasks, and wikification could be another such task. However, the pipeline architecture often suffers from error propagation.

Indeed, our motivation for Japanese wikification comes not only from the above-mentioned applications but from the lexical bottleneck problem in word segmentation. While decades of research in Japanese word segmentation has achieved a reasonable level of accuracy in benchmark corpora (Kudo et al., 2004), it is still prone to errors when it comes to real-world texts. One of the key problems is so-called unknown words, or the lack of lexical knowledge. In essence, in order to reliably identify words in text, the segmenter needs to know them in advance, by representing them directly as dictionary items (Kurohashi et al., 1994; Asahara and Matsumoto, 2000; Kudo et al., 2004) or indirectly as corpus-induced features (Neubig et al., 2011). Although lexical knowledge has been maintained by experts, it is impractical to adapt it to web-scale texts by hand (Murawaki and Kurohashi, 2008). Recognizing product names, for example, is important for downstream applications, but they are too large in number and undergo rapid transition. In light of this, one might think that Wikipedia is a promising source of lexical knowledge.

However, Wikipedia, as it is, cannot be used as a word dictionary due to segmentation mismatch. Every designer of the word segmentation task needs to define her own segmentation standard because it is left unspecified by Japanese orthography. External lexical resources like Wikipedia are full of compounds and thus incompatible with the standard. What is worse, the rise of corpus linguistics creates demand for consistency of the notion of word, with which the NLP community has been loose (Maekawa et al., 2014). This results in a book-length list of segmentation rules that often involve semantics (Ogura et al., 2011). Thus it is virtually impossible to automatically adjust segmentation of external resources.

One stop-gap solution is to ignore existing segmentation standards when incorporating external resources. In fact, this is the way a "word" dictionary named NEologd[1] was semi-automatically constructed by an industrial company in 2015. On the one hand, the intentional violation of segmentation standards makes quantitative evaluation-driven improvement difficult. On the other hand, it captures a real picture of downstream applications. In many cases, inconsistent segmentation is acceptable as long as *entities*, which are often represented by compounds, can be extracted from segmented texts. A question then is how to incorporate external resources without sacrificing quantitative evaluation.

Here, we propose to define a closely related but separate task that directly links given texts to an external resource. By doing so, we can focus on what applications want, without being bothered by segmentation mismatch. In the case of Wikipedia, a natural solution is wikification. Of course, simply defining another task does not eliminate the problem, but we believe that the disentanglement at the level of task design makes it easier to develop an elegant solution.

Another advantage over the direct incorporation of Wikipedia into word segmentation is that wikification covers *disambiguation* in addition to detection since entities sharing the same name are disambiguated into unique pages in Wikipedia. Although disambiguation is useful for appli-

[1] https://github.com/neologd

(a)

(b)

```
[[尾道市|尾道]]は[[西瀬戸自動車道|しまなみ海道]]の入り口でもあります。
```

(c)

gloss	ID	PID	word	POS	wikification
Onomichi	001	002	尾道	名詞	B 尾道市
TOP	002	009	は	助詞	
Shimanami	003	004	しまなみ	名詞	B 西瀬戸自動車道
Kaidō	004	005	海道	名詞	I
GEN	005	006	の	助詞	
entrance	006	007	入り口	名詞	
COP	007	008	で	助動詞	
too	008	009	も	助詞	
be	009	010	あ	動詞	
SUF	010	011	り	語尾	
POL	011	012	ま	助動詞	
SUF	012	013	す	語尾	
.	013	-1	。	補助記号	

Figure 1: An example of wikification. (a) Mappings from mentions to entities. (b) Wikitext as in Wikipedia. (c) Our annotation.

cations, it is beyond the scope of word segmentation. Thus when Wikipedia is converted into a word dictionary, we have to waste effort merging entities with the same name. By contrast, wikification enables us to fully exploit information provided by Wikipedia.

With this idea, we first design the task of Japanese wikification. We then annotate multiple genres of text for wikification. Finally we test the new corpora with a simple wikifier, which will hopefully be used as a baseline in future research. The new corpora also contain gold-standard word segmentation (and part-of-speech (POS) and dependency relations for their main parts), paving the way for jointly modeling these tasks.

2. Corpus Annotation

2.1. Preliminaries

Wikification consists of two subtasks, *mention detection* and *entity disambiguation*. Since a large portion of previous studies has only focused on the latter, we specifically call the combination of the two subtasks *end-to-end linking*.

Mention detection is the task of identifying consecutive sequences of characters in text that can link to Wikipedia articles. Such a character sequence is called a *mention*. For example, Figure 1(a) contains two mentions, "尾道" (*Onomichi*, a place name) and "しまなみ海道" (*Shimanami Kaidō*, a common alias for an expressway).

Entity disambiguation is the task of linking each mention to a Wikipedia article, which is called an *entity*. The title of a Wikipedia article is used as the entity name because each Wikipedia article has a unique title. In the above example, "尾道" (*Onomichi*) is linked to an entity named "尾道市" (*Onomichi City*). Similarly, "しまなみ海道" (*Shimanami Kaidō*) is linked to "西瀬戸自動車道" (*Nishiseto Expressway*, the official name) although they do not resemble each other. Wikification serves as a normalizer of surface forms.

In addition to variability, ambiguity is also addressed in wikification since the mappings from mentions to entities are many-to-many mappings. Depending on context, mention "尾道" (*Onomichi*) can refer to entity "尾道駅" (*Onomichi Station*) too.

As illustrated in Figure 1(b), Wikipedia's articles are written in *wikitext*. They are full of mention-to-entity mappings from which we might be able to train a wikifier. However, we decide to annotate corpora by ourselves primarily because Wikipedia is too inconsistent to be used directly as training data. As described in Wikipedia's manual of style, only the first occurrence of an entity is usually linked, and entities understood by most readers are left unlinked.[2]

2.2. Sources

The main corpora we work on are subsets of the Balanced Corpus of Contemporary Written Japanese (BCCWJ) (Maekawa et al., 2014). In BCCWJ, sentences are segmented into words and each word is annotated with a POS tag and pronunciation. In addition, Mori et al. (2014) augmented them with gold-standard dependency relations.

Figure 1(c) depicts our annotation for wikification. The sentence is segmented into POS-tagged words. Each word is given an ID and its parent word ID (PID). What we add is marked with the dotted rounded rectangle. We adopt the BIO labeling scheme, where B, I, and O stand for beginning, inside, and outside, respectively. For efficiency, O is not marked. A B and zero or more successive I's represent a mention while a character sequence right after the B tag indicates a corresponding entity. We assume that a mention does not subdivide a word. We observe that the assumption holds well since words we use are so-called *short-unit words* as specified by Ogura et al. (2011). One type of exceptions is compounds formed by fusing two words. For example, "送配電" (*electric transmission and distribution*) is a fusion of two words "送電" (*electric transmission*) and "配電" (*electric distribution*), both of which are valid mentions. Since "送配電" is considered a word, we have no way of tagging these sub-word mentions.

The sources of BCCWJ range from newspaper articles to blogs. We select subsets of two subcorpora: (1) white papers (OW) and (2) Yahoo! Blog (OY). We also create a pilot corpus of Twitter short messages, or *tweets*. It has gold-standard word segmentation but no dependency relations. We still use the format shown in Figure 1(c), and the dummy PID -1 is attached. Although previous studies on tweet wikification exploit Twitter-specific features such as hashtags and user timelines (Cassidy et al., 2012; Huang et al., 2014), we leave their incorporation for future work.

[2] https://en.wikipedia.org/wiki/Wikipedia:Manual_of_Style/Linking#Overlinking_and_underlinking. Japanese Wikipedia has roughly the same set of policies.

Corpus	#Docs	#Sents	#Words	#Chars	#Mentions	#Chars/mention	#Training sents	#Test sents
BCCWJ-OW	8	504	23,952	34,204	3,486	2.83	343	161
BCCWJ-OY	34	509	9,239	13,341	922	2.90	310	199
Twitter	NA	2,942	37,009	58,446	3,649	3.10	2442	500

Table 1: Corpus specifications.

Corpus	Mention detection	End-to-end linking
BCCWJ-OW	89.5%	86.1%
BCCWJ-OY	87.8%	83.2%
Twitter	76.8%	73.9%

Table 2: Annotation quality (F1 measures).

2.3. Annotation Process

We began by choosing a snapshot of Japanese Wikipedia. We used an XML dump dated 12 May 2015.[3] For the sake of reproducibility, we decided to stick with this version. For practical use, however, it may be desirable to provide another set of annotated corpora that are kept up-to-date with the most recent snapshot.

Next, we collected mention-to-entity mappings from Wikipedia. Each mention candidate m is mapped to a set of potential entities $\{e_1, \cdots, e_k\}$. Specifically, we used the following resources.

- Article title (mention identical with the entity name).
- Boldface text in the lead sentence of an article, which is sometimes not identical with the article title. For example the article for "ウラジーミル・プーチン" (Vladimir Putin) introduces him as "ウラジーミル・ウラジーミロヴィチ・プーチン" (Vladimir Vladimirovich Putin).
- The `wrongtitle` template, which indicates that C#, for example, has the article title of "C Sharp" because of technical restrictions.
- Title of a redirect, or contentless page that points to another article. For example, "尾道" (*Onomichi*) is redirected to "尾道市" (*Onomichi City*).
- Links in wikitext, as illustrated in Figure 1(b).

Note that potential entities were limited to valid Wikipedia articles. We ignored links to titles with which Wikipedia did not have articles (i.e., no NIL detection or clustering). We also excluded disambiguation pages, which simply listed possible referents. As a result, we obtained 2.5 million unique mentions for 909 thousand entities. On average, a mention is mapped into 1.09 entities while each entity has 2.98 unique mentions.

The automatically generated mention-to-entity mappings were used to enumerate possible mention-entity pairs for a given sentence in a target sentence. To do this efficiently, we built a trie data structure of mention candidates. For a given sentence, mention candidates were first enumerated using the trie, and then each of the candidates was mapped to potential entities. Since the naïve enumeration gave too many candidates, we applied a simple filtering rule: a mention candidate whose pre-computed *LinkProb* was below a threshold was dropped. *LinkProb* will be defined later in Table 3.

We asked annotators to choose the best mention-entity pair from the candidate list (or discard all if appropriate). They were allowed to add mention-entity pairs not in the list since the automatically-generated mappings were by no means perfect.

2.4. Annotation Guidelines

Ling et al. (2015) pointed out that lack of annotation guidelines posed a serious problem in the task of wikification. To keep annotation consistent, we adopt the following policies.

Any topics. Unlike many previous studies, we do not limit the scope of wikification to named entities. All mentions of both common and proper noun phrases are to be tagged as long as they have corresponding Wikipedia articles.

Exhaustive annotation. While several previous studies (Mihalcea and Csomai, 2007; Guo et al., 2013), as well as Wikipedia itself, focus on "important" concepts, we tag as many mentions as possible. It is impossible to determine importance in an objective way. Note again that unlike in Kulkarni et al. (2009), no NILs are covered.

Prefer longer mentions. Mention "日本国" (*State of Japan*), for example, is to be linked to entity "日本" (*Japan*) and should not be divided into "日本" (*Japan*, entity identical with the mention) and "国" (*state*, mapped to a synonymous entity "国家"). Sometimes a pair of mention candidates do not nest within one another but simply overlap. For example, "舞台芸術作品" (*performing art work*) has two candidate mentions "舞台芸術" (*performing art*) and "芸術作品" (*art work*). We select the former considering the internal structure of the noun phrase "[[舞台 芸術] 作品]".

Allow topical matching. Because Wikipedia is not a carefully designed ontology but an encyclopedia for human readers, it only has a single article for entities with substantially overlapping content. Such entity pairs include hyponyms, a predecessor of an organization, and other topically related entities such as "法案" (*bill*) and "法律" (*law*), and "メーカー" (*manufacturer*) and "製造業" (*manufacturing*). We cover this kind of mention-entity pairs. Although this makes wikification too ill-defined to be an entity linking task, we believe that as long as we rely on Wikipedia, loose topical matching is unavoidable.

2.5. Corpus Statistics

The results of our annotation are summarized in Table 1. Table 2 indicates the qualities of annotation. We asked one annotator to independently annotate sampled data (294 sentences in total), and the results were compared against the

[3] `http://dumps.wikimedia.org/jawiki/`

	Feature	Description
	Bias	Always active.
Mention	IDF$_f$	$\log(\|E\|/df_f(m))$, where $f \in \{\text{title}, \text{content}, \text{link}\}$, E is the set of entities and $df_f(m)$ is # of entities containing m in its field f.
	Keyphraseness	$df_{\text{link}}(m)/df_{\text{content}}(m)$.
	LinkProb	$tf_{\text{link}}(m)/tf_{\text{content}}(m)$.
	SNIL	# of entities that are identical with m.
	SNCL	# of entities that contain m.
	MLen	Bins for # of characters in m.
	CType$_i$	Feature template for the character types (Hiragana, Katakana, Latin, Punctuation, etc.) of m. $i \in \{1, -1, 2, 4\}$, where each value stands for the first character, the last character, the first and last characters, and the first two and last two characters, respectively.
	SEEWIKT	m is associated with the `See wiktionary` template (indicating content that differs from the general meaning).
	LINKWIKT	m is associated with the `Wiktionary` template.
Entity	Inlinks	# of articles linking to e.
	Outlinks	# of articles linking from e.
	Redirects	# of redirect pages linking to e.
	ELen	Bins for # of characters in the title of e.
	WikiStats	# of times e was visited in a year.
Pair	PriorProb	$tf_{\text{link}}(m, e)/\sum_e tf_{\text{link}}(m, e)$, where $tf_f(m, e)$ is # of times m linking to e.
	TF$_f$	$tf_f(m, e)/\|f\|$, where $f \in \{\text{title}, \text{link}, \text{first sentence}, \text{first paragraph}, \text{content}\}$.
	NCT	Whether m contains the title of e.
	TCN	Whether the title of e contains m.
	TEN	Whether the title of e is equal to m.

Table 3: Features for the wikifier.

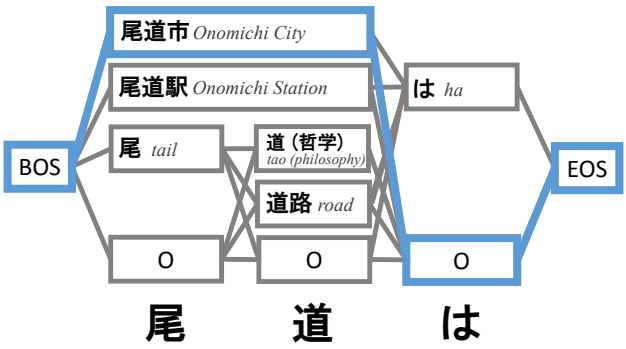


Figure 2: A lattice of mention-entity pair candidates. O is a special node for the outside of any mention. The selected path is indicated by bold lines.

gold-standard corpora. The qualities were measured using the F1 measures (see Section 3.1. for evaluation measures). They varied inversely with the formality of text genre. The white papers (BCCWJ-OW) were the easiest, the microblogs (Twitter) were the most difficult, and blogs (BCCWJ-OY) fell in between. It was not easy for outsiders to understand high-context communication in Twitter.

3. Experiments

To test the new corpora, we implemented a simple wikifier and investigated its performance. We hope that this wikifier will be used as a baseline in future research.

3.1. Settings

The wikifier was evaluated with respect to (1) mention detection and (2) end-to-end linking, each of which was measured with the standard recall, precision, and F1 measures. We compared two input formats: (1) raw sentences and (2) sentences with gold-standard word segmentation. For the latter, the wikifier only considered mention candidates whose boundaries coincided with word boundaries, both in training and test stages.

We separately assessed three corpora. Each of them was split into training and test data.

3.2. Wikifier

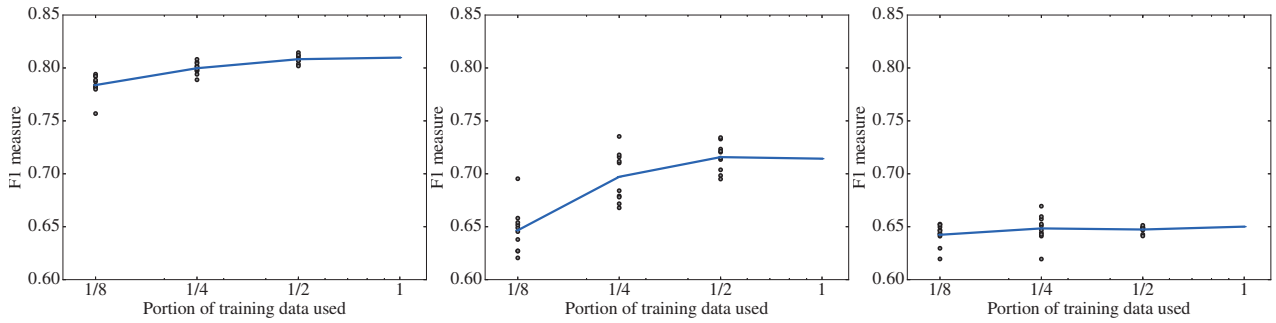
Given a sentence, the wikifier first enumerates possible mention-entity pairs (m, e) using mappings described in Section 2.3. As shown in Figure 2, candidates are organized into a lattice so that non-overlapping mentions can be selected.

A score is assigned to each node, and the Viterbi algorithm is used to find the path that maximizes the cumulative score. Each node score is the inner product of the feature vector $\phi(m, e)$ and the corresponding weight vector. Table 3 shows features we use. They are mostly based on previous studies (Meij et al., 2012; Guo et al., 2013; Huang et al., 2014), but we add several features for the character-based wikifier. Note that the wikifier behaves like a most-frequent sense baseline because currently no context feature is used for disambiguation.

We define a loss function and performed max-margin training. The model was trained with stochastic gradient descent using the adaptive Adam algorithm (Kingma and Ba, 2014)

Corpus	Input	Mention detection			End-to-end linking		
		Recall	Precision	F1	Recall	Precision	F1
BCCWJ-OW	Raw sentences	78.7%	82.1%	80.3%	77.9%	81.3%	79.6%
	Gold-standard segmentation	80.2%	83.3%	81.7%	79.4%	82.6%	81.0%
BCCWJ-OY	Raw sentences	76.7%	81.3%	78.9%	69.4%	73.5%	71.4%
	Gold-standard segmentation	79.2%	81.7%	80.4%	72.9%	75.3%	74.1%
Twitter	Raw sentences	67.7%	75.3%	71.3%	61.7%	68.7%	65.0%
	Gold-standard segmentation	73.9%	78.0%	75.9%	67.2%	71.0%	69.1%

Table 4: Performance of the wikifier.



(a) BCCWJ-OW (b) BCCWJ-OY (c) Twitter

Figure 3: Results of end-to-end linking with varying sizes of training data. The gray dots represent independent runs (10 for each set) while the blue lines denote their averages.

with L2 regularization. The number of iterations was 20, the minibatch size was 5, and weight averaging was performed.

3.3. Results

Table 4 shows the performance of the wikifier. Again, it correlated negatively with the formality of text genre. For the formal texts (BCCWJ-OW), even the simple model worked reasonably well.

Gold-standard segmentation consistently increased scores as expected. The gaps between the two types of input were larger for Twitter.

We can confirm an observation by Guo et al. (2013) that mention detection is more challenging than entity disambiguation. The drops in score from mention detection to end-to-end linking were relatively small even though the current features did not consider context at all for disambiguation.

Figure 3 shows the F1 measure of end-to-end linking as a function of the size of training data. Except for BCCWJ-OY, the size of training data did not strongly affect the wikifier's performance probably because it employed no lexicalized feature.

4. Discussion

As expected, the overwhelming majority of mentions are nouns and noun compounds. In BCCWJ, roughly half of nouns (words) are covered by mentions although the number is unfairly lowered by some grammaticalized nouns. We observe that Wikipedia is characterized by a low and imbalanced coverage of too basic terms. It appears difficult for volunteers to write encyclopedia entries for them. For polysemous terms, specialized meanings are overrepresented in Wikipedia. For example, "声明" means *statement* in ordinary speech but Wikipedia only has an article for *Buddhist chant*. High coverage of nouns and noun compounds might be achieved with an amalgam of the encyclopedia and a dictionary.

Although we focused on Japanese, we believe that the key ideas presented in this paper are applicable to other languages with *scriptiones continuae*. It is well known that Chinese also suffers from the problematic notion of word (Xia, 2000). While the Chinese language community has explored character-level analysis (Zhao, 2009), it is likely that longer units like mentions are also useful. Word segmentation and wikification might help each other because mention boundaries coincide with word boundaries. Similarly, the presence of a mention indicates that only its head word is modified from outside. A corpus with the multiple layers of annotation like ours enables us to explore this research direction.

5. Conclusion

In this paper, we reported the details of annotated corpora for Japanese wikification. We specified annotation policies, annotated corpora for wikification, and tested them with a baseline wikifier.

The main motivation behind our study is to design an elegant framework of Japanese language analysis in the presence of heterogeneous lexical resources. In the future, we plan to jointly solve the two closely related tasks, word segmentation and wikification.

6. Acknowledgements

This work was partly supported by JSPS Grants-in-Aid for Scientific Research Grant Number 26280084.

7. Bibliographical References

Asahara, M. and Matsumoto, Y. (2000). Extended models and tools for high-performance part-of-speech tagger. In *Proc. of COLING 2000*, pages 21–27.

Cassidy, T., Ji, H., Ratinov, L., Zubiaga, A., and Huang, H. (2012). Analysis and enhancement of wikification for microblogs with context expansion. In *Proc. of COLING*, pages 441–456.

Chang, M.-W., Ratinov, L., Roth, D., and Srikumar, V. (2008). Importance of semantic representation: Dataless classification. In *Proc. of AAAI*, pages 830–835.

Ferragina, P. and Scaiella, U. (2010). TAGME: On-the-fly annotation of short text fragments (by Wikipedia entities). In *Proc. of CIKM*, pages 1625–1628.

Gabrilovich, E. and Markovitch, S. (2007). Computing semantic relatedness using Wikipedia-based explicit semantic analysis. In *Proc. of IJCAI*, pages 1606–1611.

Guo, S., Chang, M.-W., and Kiciman, E. (2013). To link or not to link? a study on end-to-end tweet entity linking. In *Proc. of NAACL-HLT*, pages 1020–1030.

Huang, H., Cao, Y., Huang, X., Ji, H., and Lin, C.-Y. (2014). Collective tweet wikification based on semi-supervised graph regularization. In *Proc. of ACL*, pages 380–390.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv:1412.6980*.

Kudo, T., Yamamoto, K., and Matsumoto, Y. (2004). Applying conditional random fields to Japanese morphological analysis. In *Proc. of EMNLP*, pages 230–237.

Kulkarni, S., Singh, A., Ramakrishnan, G., and Chakrabarti, S. (2009). Collective annotation of Wikipedia entities in web text. In *Proc. of KDD*, pages 457–466.

Kurohashi, S., Nakamura, T., Matsumoto, Y., and Nagao, M. (1994). Improvements of Japanese morphological analyzer JUMAN. In *Proc. of The International Workshop on Sharable Natural Language Resources*, pages 22–38.

Ling, X., Singh, S., and Weld, D. (2015). Design challenges for entity linking. *Transactions of the Association for Computational Linguistics*, 3:315–328.

Maekawa, K., Yamazaki, M., Ogiso, T., Maruyama, T., Ogura, H., Kashino, W., Koiso, H., Yamaguchi, M., Tanaka, M., and Den, Y. (2014). Balanced corpus of contemporary written Japanese. *Language Resources and Evaluation*, 2(48):345–371.

Meij, E., Weerkamp, W., and de Rijke, M. (2012). Adding semantics to microblog posts. In *Proc. of WSDM*, pages 563–572.

Mihalcea, R. and Csomai, A. (2007). Wikify!: Linking documents to encyclopedic knowledge. In *Proc. of CIKM*, pages 233–242.

Mori, S., Ogura, H., and Sasada, T. (2014). A Japanese word dependency corpus. In *Proc. of LREC*, pages 753–758.

Murawaki, Y. and Kurohashi, S. (2008). Online acquisition of Japanese unknown morphemes using morphological constraints. In *Proc. of EMNLP 2008*, pages 429–437.

Neubig, G., Nakata, Y., and Mori, S. (2011). Pointwise prediction for robust, adaptable Japanese morphological analysis. In *Proc. of ACL-HLT*, pages 529–533.

Ogura, H., Koiso, H., Fujiike, Y., Miyauchi, S., Konishi, H., and Hara, Y., (2011). *Rules governing the morphological analysis contained in the BCCWJ, 4th ed. (part2)*. (in Japanese).

Ratinov, L. and Roth, D. (2012). Learning-based multi-sieve co-reference resolution with knowledge. In *Proc. of EMNLP-CoNLL*, pages 1234–1244.

Xia, F. (2000). The segmentation guidelines for the Penn Chinese Treebank (3.0). Technical report, University of Pennsylvania Institute for Research in Cognitive Science.

Zhao, H. (2009). Character-level dependencies in Chinese: Usefulness and learning. In *Proc. of EACL*, pages 879–887.