A Japanese Chess Commentary Corpus

Shinsuke Mori[1] **John Richardson**[2] **Atsushi Ushiku**[3]
Tetsuro Sasada[4] **Hirotaka Kameko**[5] **Yoshimasa Tsuruoka**[6]
[1,4]Academic Center for Computing and Media Studies, Kyoto University
[2,3]Graduate School of Informatics, Kyoto University
[5,6]Graduate School of Engineering, The University of Tokyo
[1,2,3,4]Yoshidahonmachi, Sakyo-ku, Kyoto, Japan
[5,6]Hongo, Bunkyo-ku, Tokyo, Japan
[1]forest@i.kyoto-u.ac.jp, [2]john@nlp.ist.i.kyoto-u.ac.jp, [3]ushiku@ar.media.kyoto-u.ac.jp,
[4]sasada@ar.media.kyoto-u.ac.jp, [5]kameko@logos.t.u-tokyo.ac.jp, [6]tsuruoka@logos.t.u-tokyo.ac.jp

Abstract

In recent years there has been a surge of interest in the natural language prosessing related to the real world, such as symbol grounding, language generation, and nonlinguistic data search by natural language queries. In order to concentrate on language ambiguities, we propose to use a well-defined "real world," that is game states. We built a corpus consisting of pairs of sentences and a game state. The game we focus on is *shogi* (Japanese chess). We collected 742,286 commentary sentences in Japanese. They are spontaneously generated contrary to natural language annotations in many image datasets provided by human workers on Amazon Mechanical Turk. We defined domain specific named entities and we segmented 2,508 sentences into words manually and annotated each word with a named entity tag. We describe a detailed definition of named entities and show some statistics of our game commentary corpus. We also show the results of the experiments of word segmentation and named entity recognition. The accuracies are as high as those on general domain texts indicating that we are ready to tackle various new problems related to the real world.

Keywords: game commentary, named entity, symbol grounding

1. Introduction

In recent years there has been a surge of interest in the generation of natural language annotations to describe digital recordings of the real world. A notable example is sentence generation from images (Ushiku et al., 2011; Yang et al., 2011). In such studies the natural language annotations are often provided by human workers on Amazon Mechanical Turk, and thus are often somewhat artificial. Since Hashimoto et al. (2014) recorded cooking videos of recipes spontaneously posted to an Internet site (Mori et al., 2014b), there have been many other image/video datasets (Ferraro et al., 2015) published. These attempts at connecting language expressions to real world objects such as images are often called *symbol grounding* (Harnad, 1990), an exciting new area in natural language processing.

However, images, videos, and many other forms of media have ambiguities that make symbol grounding difficult. In this task we propose to use a well-defined "real world," that is game states, to concentrate on language ambiguities. The game we focus on is *shogi* (Japanese chess) (Leggett, 2009). We collected 742,286 commentary sentences in Japanese and then defined domain-specific named entities (NEs), similar to previous work on bio-medical NEs (Settles, 2004; Tateisi et al., 2002) or recipe NEs (Mori et al., 2014b). For example, "central rook" is a *shogi* strategy expression similar to "king's gambit" in chess. We finally annotated NEs for 2,508 sentences to form our game commentary corpus, which has the following distinguishing characteristics:

- The commentaries are spontaneously given by professional players or writers.
- Each commentary has a corresponding game state in a real match.
- The game states do not have any ambiguity.

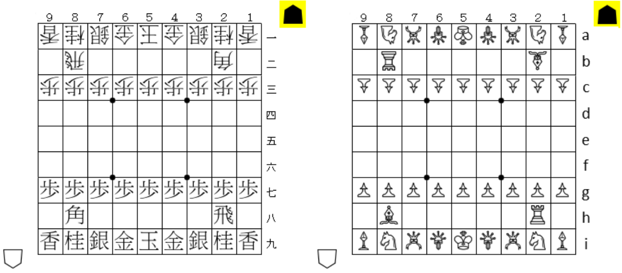
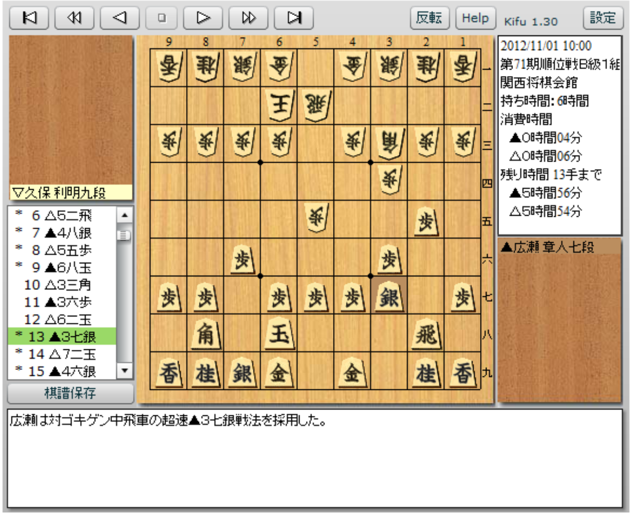


Figure 1: Starting setup of *shogi* (left: normal depiction, right: chess-like depiction).

Typical usages of our corpus include automatic commentary generation, detection of domain specific expressions, and their classification referring to game states (in the real world). There has in the past been an attempt at predicting characteristic words given a game state to generate a commentary (Kameko et al., 2015b). In this study, however, they only predict words (e.g. "king's") and do not identify concepts (e.g. "king's gambit"), nor concept types (e.g. strategy). With our corpus we can try generation using automatically generated templates (Reiter, 1995; Mori et al., 2014a) or deep learning with our NEs in place of dialog acts (Wen et al., 2015).

2. Game and Commentary

In this section we briefly explain *shogi* and its commentaries. For detailed explaination of *shogi*, please refer to (Leggett, 2009).



2012/11/01 10:00
第71期順位戦B級1組
関西将棋会館
持ち時間: 6時間
消費時間
▲0時間04分
△0時間06分
残り時間 13手まで
▲5時間56分
△5時間54分

▽久保 利明九段

▲広瀬 章人七段

* 6 △5二飛
* 7 ▲4八銀
* 8 △5五歩
* 9 ▲6八玉
10 △3三角
11 ▲3六歩
12 △6二玉
* 13 ▲3七銀
* 14 △7二玉
* 15 ▲4六銀

広瀬は対ゴキゲン中飛車の超速▲3七銀戦法を採用した。

Figure 2: A game state and a commentary broadcast via Internet. The commentary by another professional is placed below the board, which says "Hirose$_{\mathrm{Hu}}$ adopted$_{\mathrm{Ac}}$ super quick S-3g strategy$_{\mathrm{St}}$ against cheerful central rook$_{\mathrm{St}}$."

2.1. Shogi

Shogi is a two-player board game similar to chess. The goal of each player is to capture the opponent's king. Each player moves one piece alternately. Figure 1 shows the starting setup. In addition to six kinds of chess-like pieces with similar moves, *shogi* has three more kinds of pieces: gold (e.g. the piece at 4a), silver (e.g. at 3a), and lance (e.g. at 1a). The biggest difference from chess is that a player can drop a captured piece back onto the board instead of making a move. Another major difference is that almost all pieces, except for golds and kings, can promote if they move to, or move from the opponet's territory (the furthest three ranks).

2.2. Shogi Commentary

Professional players and writers give commentaries of professional matches for *shogi* fans explaining the reasoning behind moves, evaluation of the current state, and suggest probable next moves. These commentaries are broadcast with the corresponding board state as shown in Figure 2. We thus have many pairs of board states and commentary sentences in natural language. Those sentences are almost grammatically correct. We checked randomly selected 100 sentences and found no typo nor grammatical error. Among them 12 sentences are actually a phrase like "a practical move." The omitted subject may be "this game" or "this move."

These commentaries mostly concern the game itself inluding notations for specifying moves, such as "△１四香とすれば決戦" (white's Lx1d would shift the phase to the end game), but sometimes the commentaries also include information irrelevant to the board state, such as information about the players.

3. Shogi Commentary Expressions

Commentaries contain many domain specific expressions (words or multi-word expressions), which can be categorized into groups in a similar way to the general domain or bio-medical NEs. All players are familiar with these expressions (e.g. King's gambit, Ruy Lopez, etc. in chess) and category names (e.g. opening).
To facilitate various studies, we created detailed definitions of *shogi* NEs and annotated NEs for thousands of sentences. In this section, we describe the annotation standard.

3.1. Move Notation

Shogi has a well-defined notation to record games. The notation of a move is decomposed into the following components. These categories are basically finite, but we include misspelled expressions as well.

Tu: Expressions indicating the turn. This category only contains "先手" (black), "後手" (white), "▲" (black), and "△" (white).

Po: Positions denoted by two numerals (one Arabic numeral for file and one Chinese numeral for rank).

Pi: Piece names including promoted ones (14 types).

Ps: Piece specifiers. When there are multiple pieces of the same type which can be moved to the specified position, we add an expression specifying the original position such as "右" (right) indicating that the right one was moved.

Mc: Move compliment. There are only two expressions: "成" (promoted) and "不成" (non-promoted).

All move notations match the following pattern:

$$\mathrm{Tu\ Po\ Pi\ (Ps)\ (Mc)},$$

where parentheses mean that the content is optional.

3.2. Move Descriptions

For some moves, a commentator explains their meaning using the following expressions:

Mn: Move name such as "王手" (check).

Me: Move evaluation such as "好手" (good move).

3.3. Opening Expressions

Opening sequences have set names, which appear frequently.

St: Strategy names. As with chess, *shogi* has many attacking formations with various names. This class is almost closed, but sometimes new openings are invented. An example is "ゴキゲン中飛車" (cheerful central rook).

Ca: Castle names. Defensive formations also have names. This class is also almost closed with some exceptions like "ミレニアム" (Millenium formation), which arose in the year of 2000.

These may be the first target for the symbol grounding research, an application of our corpus.

3.4. Position Evaluation

The most important commentaries are those concerning evaluation of the current board state, for example “black is winning.” The class for this type of commentary includes adjectival expressions and simple sentences consisting of a subject and a predicate with arguments.

Ev: Evaluation expressions about the entire board. This category does not include those from a specific viewpoint covered by the followings.

Ee: Other evaluation expressions focusing on a certain aspect. Examples are “駒得” (gaining pieces) and “配置が良い” (pieces are well positioned).

3.5. Expressions for Description of Board Positions

Commentators use the following expressions to describe board states.

Re: Region on the board, such as “中央” (center), “４筋” (4th file), and “３段目” (3rd rank).

Ph: Phase of the match, such as “序盤” (opening), “中盤” (middlegame), and “終盤” (endgame), including vague ones such as “終盤の入り口” (start of endgame).

Pa: Piece attributes. Every piece has its own movement and commentators use special expressions for it. For example, “道” (path) is used to denote bishop’s diagonal lines and rook’s orthogonal lines. There are special expressions to denote relative positions of a piece like “腹” (belly) meaning the side squares of a piece.

Pq: Piece quantity. Usually it is a pair of a number and a counter word. This also includes expressions such as “切れ” (lack of) and “豊富” (abundant).

3.6. Describing Events Outside the Board

Commentators sometimes refer to issues outside of the board but related to the match. They can be classified into the followings types:

Hu: Names of players, commentators, etc. including their title, such as “名人” (champion). This category also contains expressions for groups of players and places such as “検討室” (discussion room) which behaves like a human. Names in expressions belonging to other types are excluded like Ishida style$_{\text{St}}$.

Ti: Expressions for the total time spent, the time spent on the current move and the time remaining. In addition to concrete expressions, like “10 minutes,” this includes abstract ones such as “長時間” (long time).

3.7. Actions

Unlike the general NE definitions, we decided to incorporate verbal expressions including copula verbs followed by an adjective. These include passive forms and causative forms.

Ac: Verbs whose subject is a player. The action must be related to the board, such as “捨てる” (sacrifice). Thus this does not include other player actions like “close eyes.”

Ap: Verbs whose subject is a piece. For example “下がった” (retreated).

Ao: Other verbs. For example “始まる” (start), with the subject “戦い” (battle).

Note that inflectional endings are excluded in order that we can adopt a statistical method for selecting the correct inflectional ending given a *shogi* NE in sentence generation (Mori et al., 2014a).

3.8. Others

We also have a class to mark other expressions related to *shogi*.

Ot: Other important notions for *shogi*. Typical ones are noun phrases denoting the above categories themselves like “戦型” (strategy). Note that this in not included in St.

4. Game Commentary Corpus

In this section we briefly explain our annotation framework, show some statistics for our corpus, and describe the corpus availability.

4.1. Annotation Framework

As the notation for *shogi* NEs, we adopt the BIO tag system (Sang and Meulder, 2003). B, I, and O stand for begin, intermediate, and other, respectively. Each word is annotated with O, indicating that the word is not any NE, or a combination of BI tag and an NE type tag such as Hu-B, which indicates that the word is the beginning (B) of a player name (Hu). The following is the correct annotation for the commentary in Figure 2.

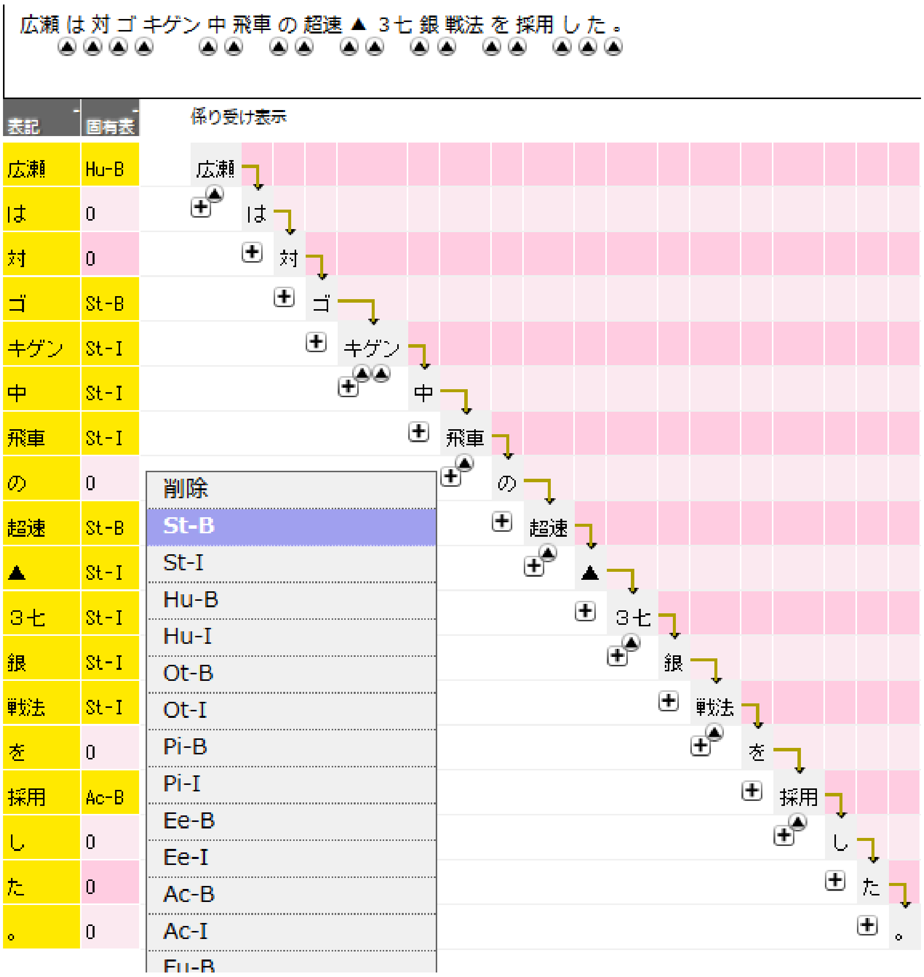
> 広瀬/Hu-B は/O 対/O ゴ/St-B キゲン/St-I 中/St-I 飛車/St-I の/O 超速/St-B ▲/St-I ３七/St-I 銀/St-I 戦法/St-I を/O 採用/Ac し/O た/O 。/O

In the BIO system, there are $2J + 1$ tags, where J is the number of the NE types. For *shogi* NE we defined $J = 21$ types and the annotation work is to choose one among 43 (= $2 \times 21 + 1$) tags for each word.

We prepared an annotation tool shown in Figure 3. We first segmented sentences automatically with a tool KyTea[1] (Neubig and Mori, 2010; Neubig et al., 2011), trained on the general domain corpus, BCCWJ (Maekawa et al., 2010) and a dictionary, UniDic (Den et al., 2008) containing 212,900 words. We then supplied the results to the tool. Finally an annotator corrected word boundaries and added BIO tags for words using the tool shown in Figure 3.

Pushing a “＋” button connects the words to form a single word and pushing a “▲” button separates a word into two words. A BIO tag is annotated to each word by selecting one among those in the pull-down menu.

[1] `http://www.phontron.com/kytea/` (accessed on 2016 Feb. 19).



広瀬 は 対 ゴ キゲン 中 飛車 の 超速 ▲ ３七 銀 戦法 を 採用 し た 。

表記	固有表
広瀬	Hu-B
は	O
対	O
ゴ	St-B
キゲン	St-I
中	St-I
飛車	St-I
の	O
超速	St-B
▲	St-I
３七	St-I
銀	St-I
戦法	St-I
を	O
採用	Ac-B
し	O
た	O
。	O

Figure 3: Annotation tool for word segmentation and BIO tags (the depenency part is not used).

Type	#Matches	#States	#Sent.	#NEs	#Words	#Char.
manu.	9	542	2,508	10,287	34,186	46,059
auto.	6,514	273,303	742,286	*3,279,851	*11,049,485	15,021,152

We counted only the states with commentaries. The values with "*" mark are based on the automatic processing results.

Table 1: Specifications of our *shogi* commentary corpus.

4.2. Statistics on the Corpus

Table 1 shows statistics of our corpus. For the number of states we counted only those with commentaries. The average number of states per match was 111. Thus 37.8% $(= 273303/(6514 \times 111))$ of the states have commentaries. From the table, for example, we can say that a manually annotated sentence contains about 3.24 NEs.

The average length of an NE is 1.27 words and 1.88 characters.

4.3. Corpus Availability

Below we briefly describe the availability of our corpus. For detailed explanations readers may visit the website `http://plata.ar.media.kyoto-u.ac.jp/data/game/home-e.html`.

The game records and the commentary sentences are distributed in the website `http://www.meijinsen.jp` (in Japanese) for a fee. Our script, available at `https://github.com/hkmk/shogi-comment-tools`, downloads the game records and the commentary sentences. The word segmentation and NE annotations are available at `http://plata.ar.media.kyoto-u.ac.jp/data/game/home-e.html`.

Game records are in KIF format. KIF files are comprised of a header and a sequence of moves. Each move is denoted as a sequence of the player, position, and piece. If it is ambiguous, a piece specifier and/or move compliment are added (see Subsection 3.1.). Some moves, thus the resulting states, have one or more commentary sentences by human experts. These commentary sentences are written below the moves, with a start delimiter "*." Our annotations are for these sentences.

5. Application

The most important applications of our corpus are text analysis such as word segmentation and *shogi* NE recognition.

Type	#Matches	#States	#Sent.	#NEs	#Words	#Char.
Test						
Shogi	2	156	731	2,365	7,161	9,470
Train						
Shogi	7	386	1,777	7,922	27,025	36,589
BCCWJ	NA	NA	57,281	—	1,339,500	1,931,751

Table 2: Corpus specifications for word segmentation and NE recognition experiments.

Training	Precision	Recall	F-measure
BCCWJ	0.872	0.907	0.889
BCCWJ + *Shogi*	0.983	0.983	0.983

Table 3: Word segmentation accuracies.

In this section we show their results and describe other potential applications.

5.1. Word Segmentation

Our corpus has word boundary information. We therefore first tested word segmentation performance. It is well known that an annotated corpus in the target domain improves the performance very well (Mori and Neubig, 2014). Thus in addition to the baseline (see Subsection 4.1.), we trained another model using our corpus additionally. Table 2 shows the experimental settings.

The results are shown in Table 3. The performance of the baseline word segmenter is very bad. Our corpus, however, improves the performance drastically. We can realize a further improvement in word segmentation on game commentaries by referring to the game states (Kameko et al., 2015a). So we can say that an accurate word segmenter for *shogi* commentaries is now available.

5.2. Named Entity Recognition

We also conducted *shogi* NE recognition. We trained a BIO2-based NE recognizer (Sasada et al., 2015) and tested it. The corpus specifications are shown in Table 2.

The precision and recall are 0.913 and 0.789, respectively. The F-measure is 0.847, which is comparable to the general domain case (around 0.9) trained from about 10,000 sentences (Sang and Meulder, 2003; McCallum and Li, 2003). These results suggest that an NE recognizer is ready to be used to detect *shogi* NEs in raw commentaries for various applications.

5.3. Symbol Grounding

One of the most interesting research directions is symbol grounding. Contrary to images or videos (Regneri et al., 2013), game states do not cause recognition problems and we can concentrate on the ambiguities on the language side. Another interesting aspect of symbol grounding to game states is that we can connect natural language expressions to computer analysis and predictions.

5.4. Others

The NE recognition and/or symbol grounding results allow for various applications. Firstly we can improve automatic commentary generation (Kaneko, 2012). Kameko et al. (2015b) proposed a method for finding characteristic words for game states and used them to generate commentaries automatically. With our corpus, one can try to use NEs instead of words to generate better commentaries.

We can also create a system for game state search by natural language queries. There is a database storing all professional matches. At present we can search for states by piece positions (Ganguly et al., 2014), but not by language expressions, such as strategy names, because there is no clear definition for them. The combination of NE recognition and symbol grounding will enable this.

Because there has never been a corpus of game commentaries related to game states, we believe that there are many other novel applications such as bilingual lexicon aquisition based on symbol grounding (Kiela et al., 2015).

6. Conclusion

In this paper, we described our *shogi* commentary corpus. The sentences in the corpus are segmented into words and annotated with domain specific NE tags. The most interesting characteristics of our corpus is that every commentary is connected to a game state (real world).

Using our corpus we have been able to obtain high quality automatic word segmentation and NE recognition. We believe this will enable NLP and AI researchers to begin to tackle various new problems such as symbol grounding, commentary generation, and intelligent game state search.

7. Acknowledgments

This work was supported by JSPS Grants-in-Aid for Scientific Research Grant Numbers 26540190 and 26280084. We are also grateful to the annotators for their contribution to the design of the guidelines and the annotation effort.

8. Bibliographical References

Den, Y., Nakamura, J., Ogiso, T., and Ogura, H. (2008). A proper approach to Japanese morphological analysis: Dictionary, model, and evaluation. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation*, pages 1019–1024.

Ferraro, F., Mostafazadeh, N., Huang, T.-H., Vanderwende, L., Devlin, J., Galley, M., and Mitchell, M. (2015). A survey of current datasets for vision and language research. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 207–213.

Ganguly, D., Leveling, J., and Jones, G. J. (2014). Retrieval of similar chess positions. In *Proceedings of the 37th annual international ACM SIGIR conference*, pages 687–696. ACM.

Harnad, S. (1990). The symbol grounding problem. *Physica D*, 42:335–346.

Hashimoto, A., Sasada, T., Yamakata, Y., Mori, S., and Minoh, M. (2014). KUSK dataset: Toward a direct understanding of recipe text and human cooking activity. In *Proceedings of the SixthInternational Workshop on Cooking and Eating Activities*.

Kameko, H., Mori, S., and Tsuruoka, Y. (2015a). Can symbol grounding improve low-level NLP? word segmentation as a case study. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2298–2303.

Kameko, H., Mori, S., and Tsuruoka, Y. (2015b). Learning a game commentary generator with grounded move expressions. In *Proceedings of the 2015 IEEE Conference on Computational Intelligence and Games*.

Kaneko, T. (2012). Real time commentary system for shogi. In *First Workshop on Games and NLP*.

Kiela, D., Vulić, I., and Clark, S. (2015). Visual bilingual lexicon induction with transferred convnet features. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 148–158.

Leggett, T. (2009). *Japanese chess : the game of shogi*. Tuttle Publishing.

Maekawa, K., Yamazaki, M., Maruyama, T., Yamaguchi, M., Ogura, H., Kashino, W., Ogiso, T., Koiso, H., and Den, Y. (2010). Design, compilation, and preliminary analyses of balanced corpus of contemporary written Japanese. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation*.

McCallum, A. and Li, W. (2003). Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In *Proceedings of the Seventh Conference on Computational Natural Language Learning*.

Mori, S. and Neubig, G. (2014). Language resource addition: Dictionary or corpus? In *Proceedings of the Nineth International Conference on Language Resources and Evaluation*, pages 1631–1636.

Mori, S., Maeta, H., Sasada, T., Yoshino, K., Hashimoto, A., Funatomi, T., and Yamakata, Y. (2014a). Flowgraph2text: Automatic sentence skeleton compilation for procedural text generation. In *Proceedings of the the Eighth International Conference on Natural Language Generation*, pages 118–122.

Mori, S., Maeta, H., Yamakata, Y., and Sasada, T. (2014b). Flow graph corpus from recipe texts. In *Proceedings of the Nineth International Conference on Language Resources and Evaluation*, pages 2370–2377.

Neubig, G. and Mori, S. (2010). Word-based partial annotation for efficient corpus construction. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation*, pages 2723–2727.

Neubig, G., Nakata, Y., and Mori, S. (2011). Pointwise prediction for robust, adaptable Japanese morphological analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, pages 529–533.

Regneri, M., Rohrbach, M., Wetzel, D., Thater, S., Schiele, B., and Pinkal, M. (2013). Grounding action descriptions in videos. *Transactions of the Association for Computational Linguistics*, 1:25–36.

Reiter, E. (1995). NLG vs. templates. In *Proceedings of the the Fifth European Workshop on Natural Language Generation*, pages 147–151.

Sang, E. F. T. K. and Meulder, F. D. (2003). Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Computational Natural Language Learning*, pages 142–147.

Sasada, T., Mori, S., Kawahara, T., and Yamakata, Y. (2015). Named entity recognizer trainable from partially annotated data. In *Proceedings of the Eleventh International Conference Pacific Association for Computational Linguistics*.

Settles, B. (2004). Biomedical named entity recognition using conditional random fields and rich feature sets. In *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications*, pages 33–38.

Tateisi, Y., Kim, J.-D., and Ohta, T. (2002). The GENIA corpus: an annotated research abstract corpus in molecular biology domain. In *Proceedings of the HLT*, pages 73–77.

Ushiku, Y., Harada, T., and Kuniyoshi, Y. (2011). Automatic sentence generation from images. In *Proceedings of the 19th Annual ACM International Conference on Multimedia*, pages 1533–1536.

Wen, T.-H., Gasic, M., Mrksic, N., Su, P.-H., Vandyke, D., and Young, S. (2015). Semantically conditioned LSTM-based natural language generation for spoken dialogue systems. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 207–213.

Yang, Y., Teo, C. L., III, H. D., and Aloimonos, Y. (2011). Corpus-guided sentence generation of natural images. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*.