A Metric for Evaluating Discourse Coherence based on Coreference Resolution

Ryu Iida[1] *Takenobu Tokunaga*[1]
(1) Tokyo Institute of Technology, W8-73, 2-12-1 Ohokayama Meguro Tokyo, 152-8552 Japan
{ryu-i,take}@cl.cs.titech.ac.jp

ABSTRACT
We propose a simple and effective metric for automatically evaluating discourse coherence of a text using the outputs of a coreference resolution model. According to the idea that a writer tends to appropriately utilise coreference relations when writing a coherent text, we introduce a metric of discourse coherence based on automatically identified coreference relations. We empirically evaluated our metric by comparing it to the entity grid modelling by Barzilay and Lapata (2008) using Japanese newspaper articles as a target data set. The results indicate that our metric better reflects discourse coherence of texts than the existing model.

KEYWORDS: discourse coherence, coreference resolution, evaluation metric.

Proceedings of COLING 2012: Posters, pages 483–494,
COLING 2012, Mumbai, December 2012.

1 Introduction

The task of automatically evaluating discourse coherence has recently received much attention (Karamanis et al., 2004; Barzilay and Lapata, 2008; Lin et al., 2011, etc.) because it is essential for several NLP applications such as generation (Soricut and Marcu, 2006), summarisation (Lapata, 2003; Okazaki et al., 2004; Bollegala et al., 2006) and automated essay scoring (Miltsakaki and Kukich, 2000; Higgins et al., 2004). Researchers in these areas have mainly been concerned with introducing the linguistic notions of cohesion or coherence addressed in discourse theories, such as Centering Theory (Grosz et al., 1995) and Rhetorical Structure Theory (Mann and Thompson, 1988), into computational models for each task, ranging from heuristic rule-based to sophisticated machine learning-based approaches.

Some of this research has relied on the occurrence of discourse entities (e.g. NPs and pronouns) to capture cohesion of a text for indirectly estimating discourse coherence. Barzilay and Lapata (2008)'s approach, for instance, models the transition of discourse entities appearing in adjacent sentences for capturing local discourse coherence, which is derived from the notion of Centering Theory. In their approach, the plausible transition of discourse entities in a coherent text is trained together with a set of incoherent texts by using a ranking SVM (Joachims, 2002), making use of a grid of each discourse entity with regard to its grammatical role, called an *entity grid* representation.

Their approach to evaluating discourse coherence is quite useful when discourse entities explicitly appear in languages such as English. In their evaluation, they reported their coherence modeling based on the entity grid representation contributes to drastically improving accuracy on the information ordering task, which is the pairwise ranking problem given a pair of coherent and incoherent texts in English. However, in languages such as Japanese and Italian, capturing the transition of discourse entities is relatively difficult due to the frequent use of ellipses. As an example of employing the entity grid model in Japanese, Yokono and Okumura (2010) directly attempted this for representing grid using typical Japanese grammatical roles (*wa* (topic), *ga* (subj), *o* / *ni* (obj / i-obj) and others). They conducted an empirical evaluation of pairwise ranking of Japanese texts, replicating the experimental settings by Barzilay and Lapata (2008). Their result shows their model achieved around 70% in accuracy, whereas the evaluation result on the English data set reaches around 90%. This difference of performance might be caused by the frequent occurrence of ellipses. In Japanese, for example, subjects in a sentence are frequently unrealised, resulting in the less frequent occurrence of adjacent discourse entities in a same coreference, which are essential for capturing the transition of discourse entities in entity grid modelling (Barzilay and Lapata, 2008).

Against this background, we propose a metric of discourse coherence, which takes into account any pair of discourse entities in a text to capture the relationship of the entities distantly appeared in a text, which cannot not be directly exploited in the entity grid approaches. In order to evaluate discourse coherence using our metric, we utilise the outputs of a coreference resolution model (especially, the reliability of each output of the model). The assumption behind it is that one tends to appropriately utilise coreference relations when writing a coherent text, i.e. the better use of coreference relations is considered as a good indicator of coherent texts.

This paper is organised as follows. Section 2 briefly reviews the previous work on automatically evaluating discourse (local) coherence. Section 3 explains the proposed metric of evaluating discourse coherence exploiting the outputs of coreference resolution and Section 4

s_1: [John] bought [iPad2] as [a gift] for [Lucy].
s_2: However, [it] has [something amiss] with [the sound system].
s_3: As a result, [he] went to [[Lucy]'s birthday party] with no [gift].

Square-bracketed words (or phrases) stand for discourse entities.

Figure 1: Coherent input example for entity computation

introduces an NP coreference resolution model employed in the metric. Section 5 reports performance of NP coreference resolution on coherent and incoherent texts in Japanese and the effectiveness of the proposed metric on the task of information ordering comparing to an existing model. Section 6 concludes the paper and discuss our future directions.

2 Related work

There has been an increase in recent work for evaluating discourse (local) coherence of a text (Barzilay and Lapata, 2008; Karamanis et al., 2004; Lin et al., 2011; Miltsakaki and Kukich, 2000; Higgins et al., 2004, etc.), which strongly relates to the cohesion of discourse entities appearing in the text from the theoretical perspective mainly based on Centering Theory (Grosz et al., 1995). For example, Karamanis et al. (2004) and Miltsakaki and Kukich (2000) proposed a metric of coherence directly utilising the transition of centers in a text, as Centering Theory does. According to the analysis by Poesio et al. (2004), Karamanis et al. (2004) define a metric based on the numbers of missing *backward-looking centers*, each of which is a discourse entity appearing in the current utterance and was realised as most salient in the previous utterance. On the other hand, Miltsakaki and Kukich (2000) focused on investigating the relationship of the coherence of a text and the transition of centers and revealed that the rough-shift transition of centers correlates to incoherence of a text.

In these studies, one of the most important work was to represent the relationship of discourse entities and their occurrences in a text based on the transition of discourse entities, which was done in a series of studies (Barzilay and Lee, 2004; Barzilay and Lapata, 2005; Lapata and Barzilay, 2005; Barzilay and Lapata, 2008). In Barzilay and Lapata (2008), the transition of discourse entities in adjacent discourse units (e.g. sentences) is formalised as an *entity grid*, which is a matrix of discourse entities and their realised grammatical roles, because a grammatical role of a discourse entity is a good indicator of its salience. For example, a given input text shown in Figure 1, consisting of the three sentences, each discourse entity is represented in the entity grid shown in Table 1. In the entity grid, each column is filled with the corresponding label (e.g. S (subject), O (object), X (others) and – (not realised)). In the grid, the local transition of entities with regard to the labels can be seen as a generalisation of the center transition discussed in a series of Centering studies (Walker et al., 1997; Grosz et al., 1995). Therefore, exploiting the transition becomes a good indicator of (local) discourse coherence. In their work, the transition of each entity was used as a feature for distinguishing a coherent text from an incoherent one.

As an extension of Barzilay and Lapata (2008), Lin et al. (2011) took into account the use of discourse relations to revise the formulation of an entity grid. They used the four types of discourse relations (Temporal, Contingency, Comparison and Expansion) defined in the Penn Discourse Treebank (PDTB) instead of grammatical roles, which are automatically acquired by the discourse parser by Lin et al. (2011). For grid representation, they calculated the tran-

	John	iPad2	gift	Lucy	sound system	birthday party
s_1	S	O	X	X	–	–
s_2	–	S	–	–	X	–
s_3	S	–	X	X	–	X

Table 1: Entity grid of the input example in Figure 1

s'_1=(s_1): [John] bought [iPad2] as [a gift] for [Lucy].
s'_2=(s_3): As a result, [he] went to [[Lucy]'s birthday party] with no [gift].
s'_3=(s_2): However, [it] has [something amiss] with [the sound system].

Figure 2: Incoherent input example for entity computation obtained by random reordering

sition probabilities of discourse entities in a text based on the PDTB-style discourse relations (e.g. P($S_i : Comp.Arg_1 \rightarrow S_{i+1} : Exp.Arg2$)), and then these probabilities are exploited as features in a ranking SVM (Joachims, 2002). Through their empirical evaluation they reported their extension of the entity grid representation contributes to improving performance on the pairwise ordering task compared to the original entity grid model.

3 A metric for evaluating coherence based on coreference resolution

As explained in Section 2, typical approaches to modeling discourse coherence have exploited the transition of discourse entities in terms of grammatical roles or discourse relations defined in PDTB. In contrast, we estimate discourse coherence by a metric relying on the outputs of an NP coreference resolution model.

For instance, from the coherent text shown in Figure 1, the corresponding incoherent text is generated by randomly reordering sentences, one of which is as shown in Figure 2. In this incoherent text, as the pronoun "it" is placed relatively far from its antecedent "iPad2" and a distractor "birthday party" is inserted between these two expressions, the interpretation of "it" is more difficult than the case of the coherent text. As a result, applying a typical coreference resolution model to coherent and incoherent texts gives rise to the difference in the number of correctly identified coreference relations. In addition, if there is no difference in terms of the number, there may be a difference in the reliability score (i.e. predicted probability outputted by a classifier) of the resolved relations. Based on these differences, we propose a metric for evaluating discourse coherence, which is calculated according to the following two steps:

1. a coreference (or anaphora) resolution model trained with annotated coherent texts is applied to a target text T.
2. the coherence score of T is calculated from the outputs of step1 by

$$\text{coherence}(T) = \frac{1}{N}\sum_{j}^{N} \text{score}_{ana}(i, j), \quad (1)$$

where T is a target text, j is a candidate anaphor appearing in T and i is the most likely candidate antecedent of j. N is the number of candidate anaphors appearing in T. The reliability score of the coreference relation of i and j, $\text{score}_{ana}(i, j)$, is the output score (e.g. predicted probability) obtained after a coreference model is applied to T in step1.

Note that the proposed metric can also be used as one of the features for the entity grid model because it is obtained from a different perspective from the entity grid (i.e. information of the discourse entity transition). In Section 5.3 we will also demonstrate the results of the entity grid model employing our metric as a feature.

4 Coreference resolution model for a coherence metric

The proposed metric introduced in Section 3 is designed for the use of any anaphora (or coreference) resolution model. In this work, we employ an NP coreference resolution model.

According to formula (1) in Section 3, calculating our metric needs a reliability score of each anaphor and candidate antecedent pair. Recent sophisticated approaches to NP coreference range from considering the transitivity of discourse entities (Denis and Baldridge, 2007) to clustering-based approaches (Cardie and Wagstaf, 1999; Cai and Strube, 2010), but these approaches aim at obtaining globally optimised scores for a set of mentions. Therefore, it is generally difficult using such models to get a reliability score for a pair of two mentions though they typically achieved better performance than simple pairwise coreference resolution models such as Soon et al. (2001) and Ng and Cardie (2002),

In the work on Japanese anaphora resolution by Iida and Poesio (2011), they employed an ILP-based approach to optimise final outputs of NP coreference resolution in Japanese and reported it achieved better performance than simple pairwise baselines. In spite of the global optimisation by ILP, their formulation can be easily reinterpreted as follows due to the best-first constraint used in their ILP formula, which is for avoiding the redundant choice of more than one candidate antecedent:

$$\text{coref}(i,j) = \frac{P(coref|i,j) + P(anaph|j)}{2} \tag{2}$$

where j is a candidate anaphor and i is the most likely candidate antecedent of j. $P(coref|i,j)$ is calculated by a simple coreference classifier such as Ng and Cardie (2002) and $P(anaph|j)$ is the score of anaphoricity of j, which is used to exclude typical non-anaphoric mentions such as pleonastic *it*. Given equation (2), their anaphora resolution model judge as anaphoric if $\text{coref}(i,j) \geq 0.5$; otherwise non-anaphoric.

In this work, we adopt the above approach to obtain $\text{score}_{ana}(i,j)$ needed in equation (1). By using $coref(i,j)$ we define $\text{score}_{ana}(i,j)$ as follows:

$$\text{score}_{ana}(i,j) = -\log(1 - \max_{i} \text{coref}(i,j)) \tag{3}$$

The feature set and detailed configuration for model creation generally follows the original work by Iida and Poesio (2011). For creating a classifier, we used MegaM [1], an implementation of the Maximum Entropy model, with default parameter settings. As an anaphoricity determination model (Iida et al., 2005), we used the selection-then-classification model, which first selects a most likely candidate antecedent i and then determines the anaphoricity of candidate anaphor j referring to the information from a pair of i and j, because Iida et al. (2005) reported their model determines anaphoricity more precisely than a simple anaphoricity model (e.g. Ng and Cardie (2002)).

[1] http://cs.utah.edu/~hal/megam/

type	#article	#sentence	#word	coreference
train	1,753	24,263	651,986	10,206
test	696	9,287	250,901	4,396

Table 2: Statistics of annotated information in NAIST text corpus

5 Empirical Evaluation

This section first evaluates performance of NP coreference resolution on coherent and incoherent texts for exploring the possible use of these results on evaluating discourse coherence; we then conduct an empirical evaluation on ranking a pair of coherent and incoherent texts by comparing our metric with the entity grid model.

5.1 Data set

For our evaluation, we used the NAIST text corpus, which consists of Japanese newspaper articles containing manually annotated NP coreference relations. Because the corpus has no explicit boundary between training and test sets, articles published from January 1st to January 11th and the editorials from January to August were used for training and articles dated January 14th to 17th and editorials dated October to December are used for testing as done by Taira et al. (2008) and Imamura et al. (2009). Table 2 summarises the statistics of annotated coreference relations in the corpus.

Because the data set contains some texts consisting of only a sentence[2], we excluded them for our evaluation of information ordering. In line with the experiments done by Barzilay and Lapata (2008), we created 20 different texts by randomly scrambling the order of the sentences in an original text, each of which is henceforth called an *incoherent text*, while the original text is called a *coherent text*. In this evaluation, we followed Barzilay and Lapata (2008)'s experimental setting, that is, the task of pairwise ordering, i.e., to detect a coherent text given a coherent and incoherent text pair.

5.2 Experiment 1: NP coreference resolution on incoherent texts

We first evaluate performance of NP coreference resolution on both coherent and incoherent texts. During the training phase, we use only coherent texts as the training instances for creating a classifier used in each model. By using only coherent texts for training, we expect that a model appropriately identifies coreference (or anaphoric) relations in coherent texts, while it is less successful in incoherent texts. Next, classifiers induced from coherent texts are applied to either coherent or incoherent texts to investigate the difference of performance on coreference resolution.

Table 3 shows the results for the recall, precision and F-score of pairwise classification on NP coreference resolution on evaluating coherent or incoherent texts, where the 'coherent' which stands for the results on coherent texts, the 'incoherent:μ' and 'incoherent:σ' which mean the averaged score of the results on incoherent texts and its standard deviation. Table 3 demonstrates that the 'coherent' obtains better performance in F-score than 'incoherent:μ' on NP coreference resolution. It indicates that the performance of NP coreference resolution strongly correlates to discourse coherence, that is, this relative difference of performance between co-

[2]In the NAIST text corpus, 213 articles in the training set and 156 articles in the test set consist of a sentence.

	Recall	Precision	F-score
coherent	0.624	0.508	**0.560**
incoherent	0.538± 0.004	0.496 ± 0.004	0.516 ± 0.004

Table 3: Results using NP coreference resolution

herent and incoherent texts is expected to lead to better discrimination on information ordering which we discuss in Section 5.3.

5.3 Experiment 2: pairwise information ordering

We next investigate the effects of the metric proposed in Section 3 for the task of pairwise information ordering comparing the results with the entity grid model.

As a baseline model, we use a model which randomly selects a text from two given texts. Alternative baselines are variants of the entity grid model; one captures the transition of discourse entities based on lexical chaining (i.e. NPs which have identical head strings are grouped as a cluster), and the other uses the outputs of a NP coreference resolution model for the entity grid representation instead of using lexical chaining. As for the coreference resolution model for obtaining the entity grid representation, we employed the original selection-then-classification model (SCM) described in Section 4 because it performed better in the final evaluation (i.e. pairwise ordering). This may be because the original SCM tends to accurately identify coreference relations in incoherent texts as well as coherent ones, and as a result those relations are considered as less noisy inputs to the entity grid model.

For the entity grid representation in Japanese, we employed the work by Yokono and Okumura (2010), which is based on Japanese case-makers (e.g. *wa* (topic), *ga* (subject), *o* (object)) to simply identify grammatical roles of discourse entities[3]. Note that we excluded the extensions of the base entity grid modeling (e.g. separating discourse entities into two classes based on the salience of each, introduced by Barzilay and Lapata (2008)) for simplification. To create a pairwise ranker based on the entity grid modelling, we used a ranking SVM (Joachims, 2002) as Barzilay and Lapata (2008) did. In this evaluation, we also compared the entity grid models using the coherence metric based on NP coreference as a feature.

The results are shown in Table 4. These results demonstrate the entity grid models and the models based on our coherence metric achieved better accuracy than the random baseline. By comparing the entity grid models with and without coreference resolution, the results show that the former outperforms the latter. It indicates Japanese NP coreference resolution is also useful for grid representation, the same as for English coreference resolution adopted in Barzilay and Lapata (2008).

Furthermore, ranking based on our metric achieved better accuracy than the entity grid models. This is because our metric has an advantage of being able to capture the coherence and incoherence resulting from the use of long-distance coreference relations, while the entity grid model focuses on the local coherence based on discourse entities appearing in the adjacent two or three sentences.

[3]In addition to the three labels (i.e. S, O and X) in the original work by Barzilay and Lapata (2008), we also use a T(opic) label to distinguish topical words from subjects done by Yokono and Okumura (2010) to capture the Japanese grammatical aspect.

	model	accuracy (%)
	random	50.0
	entity grid (−coref)	67.3
(a)	entity grid (+coref)	70.7
(b)	proposed metric	76.1
(c)	(a) + (b)	**78.2**

Table 4: Results of pairwise information ordering

Our metric utilises the appropriateness of anaphoric functions, one of characteristics of coherence which was not directly integrated in the entity grid model. Therefore, by combining them we can expect to see an improvement in accuracy. The last row in Table 4 shows the result of the entity grid model using coreference resolution integrated with our metric as a feature. As expected, the result ((c) in Table 4) obtained the best accuracy out of all the results shown in Table 4 [4]. It indicates that long-distant coreference relations are also important for evaluating discourse coherence in a text.

6 Conclusion

In this paper we proposed a metric for evaluating discourse coherence based on the outputs of a coreference resolution model to reflect the idea that a writer tends to appropriately utilise anaphoric or coreference relations when writing a coherent text. In order to investigate the effects of the proposed metric, we conducted an empirical evaluation on a pairwise ordering task, taking the NAIST text corpus as a target data set. The results of our evaluation demonstrated that the metric calculated using the outputs of NP coreference resolution achieved better accuracy than the entity grid model (Barzilay and Lapata, 2008). Moreover, the result of integrating the metric with the entity grid model shows the improvement of 7 points in accuracy.

In this work, we focused on the use of NP coreference resolution as cues for evaluating discourse coherence in a text. However, even if we refer to coreference relations as indicators of discourse coherence, the relations are sometime sparse in a text, resulting in assigning an inappropriate score to it. One simple way to avoid this problem is to take into account other types of reference behaviour, such as zero anaphora and bridging anaphora, because this type of reference function can often relate distant discourse fragments (e.g. two clauses placed far from each other). In addition, although we focused on exploiting the relationship of discourse entities in terms of anaphoric functions, the (latent) topic transition in a text is another key for capturing text coherence, as discussed by Chen et al. (2009). Therefore, one interesting issue for discourse coherence is how to integrate the above factors into existing coherence models.

References

Barzilay, R. and Lapata, M. (2005). Modeling local coherence: An entity-based approach. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics* (*ACL 2005*), pages 141–148.

Barzilay, R. and Lapata, M. (2008). Modeling local coherence: An entity-based approach. *Computational Linguistics*, 34(1):1–34.

[4] Note that the difference of (b) and (c) is statistically significant (McNemar's test, $p < 0.05$)

Barzilay, R. and Lee, L. (2004). Catching the drift: Probabilistic content models, with applications to generation and summarization. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics* (*HLT-NAACL 2004*), pages 113–120.

Bollegala, D., Okazaki, N., and Ishizuka, M. (2006). A bottom-up approach to sentence ordering for multi-document summarization. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics* (*COLING-ACL 2006*), pages 385–392.

Cai, J. and Strube, M. (2010). End-to-end coreference resolution via hypergraph partitioning. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010)*, pages 143–151.

Cardie, C. and Wagstaf, K. (1999). Noun phrase coreference as clustering. In *Proceedings of 1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pages 82–89.

Chen, H., Branavan, S. R. K., Barzilay, R., and Karger, D. R. (2009). Global models of document structure using latent permutations. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics* (*HLT-NAACL 2009*), pages 371–379.

Denis, P. and Baldridge, J. (2007). Joint determination of anaphoricity and coreference resolution using integer programming. In *Proceedings of Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics* (*HLT-NAACL 2007*), pages 236–243.

Grosz, B. J., Joshi, A. K., and Weinstein, S. (1995). Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 21(2):203–226.

Higgins, D., Burstein, J., Marcu, D., and Gentile, C. (2004). Evaluating multiple aspects of coherence in student essays. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics* (*HLT-NAACL 2004*), pages 185–192.

Iida, R., Inui, K., and Matsumoto, Y. (2005). Anaphora resolution by antecedent identification followed by anaphoricity determination. *ACM Transactions on Asian Language Information Processing* (*TALIP*), 4(4):417–434.

Iida, R. and Poesio, M. (2011). A cross-lingual ilp solution to zero anaphora resolution. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies* (*ACL-HLT 2011*), pages 804–813.

Imamura, K., Saito, K., and Izumi, T. (2009). Discriminative approach to predicate-argument structure analysis with zero-anaphora resolution. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP* (*ACL-IJCNLP 2009*), pages 85–88.

Joachims, T. (2002). Optimizing search engines using clickthrough data. In *Proceedings of the ACM Conference on Knowledge Discovery and Data Mining (KDD 2002)*, pages 133–142.

Karamanis, N., Poesio, M., Mellish, C., and Oberlander, J. (2004). Evaluating centering-based metrics of coherence using a reliably annotated corpus. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics* (*ACL 2004*), pages 391–398.

Lapata, M. (2003). Probabilistic text structuring: Experiments with sentence ordering. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics* (*ACL 2003*), pages 545–552.

Lapata, M. and Barzilay, R. (2005). Automatic evaluation of text coherence: Models and representations. In *Proceedings of 2005 International Joint Conferences on Artificial Intelligence* (*IJCAI 2005*), pages 1085–1090.

Lin, Z., Ng, H. T., and Kan, M.-Y. (2011). Automatically evaluating text coherence using discourse relations. In *Proceeding of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language TEchnologies* (*ACL-HLT 2011*), pages 997–1006.

Mann, W. C. and Thompson, S. A. (1988). Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8(3):243–281.

Miltsakaki, E. and Kukich, K. (2000). Automated evaluation of coherence in student essays. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation* (*LREC 2000*).

Ng, V. and Cardie, C. (2002). Improving machine learning approaches to coreference resolution. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* (*ACL 2002*), pages 104–111.

Ng, V. and Cardie, C. (2002). Identifying anaphoric and non-anaphoric noun phrases to improve coreference resolution. In *Proceedings of the 19th International Conference on Computational Linguistics* (*COLING 2002*), pages 730–736.

Okazaki, N., Matsuo, Y., and Ishizuka, M. (2004). Improving chronological sentence ordering by precedence relation. In *Proceedings of the 20th International Conference on Computational Linguistics* (*COLING 2004*), pages 750–756.

Poesio, M., Stevenson, R., Eugenio, B. D., and Hitzeman, J. (2004). Centering: A parametric theory and its instantiations. *Computational Linguistics*, 30(3):309–363.

Soon, W. M., Ng, H. T., and Lim, D. C. Y. (2001). A machine learning approach to coreference resolution of noun phrases. *Computational Linguistics*, 27(4):521–544.

Soricut, R. and Marcu, D. (2006). Discourse generation using utility-trained coherence models. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics* (*COLING-ACL 2006*), pages 803–810.

Taira, H., Fujita, S., and Nagata, M. (2008). A Japanese predicate argument structure analysis using decision lists. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing* (*EMNLP 2008*), pages 523–532.

Walker, M., Joshi, A. K., and (eds.), E. P. (1997). *Centering Theory in Discourse*. Oxford Univ. Press.

Yokono, H. and Okumura, M. (2010). Incorporating cohesive devices into entity grid model in evaluating local coherence of Japanese text. In *Proceedings of the 11th International Conference on Intelligent Text Processing and Computational Linguistics* (*CICLing 2010*), pages 303–314.