Causal Relation Extraction Using Cue Phrase and Lexical Pair Probabilities

Du-Seong Chang* and Key-Sun Choi

Department of Electrical Engineering & Computer Science, KORTERM, BOLA
Korea Advanced Institute of Science and Technology
373-1, Guseong-dong, Yuseong-gu, Daejeon, 306-701, Korea
dschang@world.kaist.ac.kr, kschoi@cs.kaist.ac.kr

Abstract. This work aims to extract causal relations that exist between two events expressed by noun phrases or sentences. The previous works for the causality made use of causal patterns such as causal verbs. We concentrate on the information obtained from other causal event pairs. If two event pairs share some lexical pairs and one of them is revealed to be causally related, the causal probability of another event pair tends to increase. We introduce the lexical pair probability and the cue phrase probability. These probabilities are learned from raw corpus in unsupervised manner. With these probabilities and the Naive Bayes classifier, we try to resolve the causal relation extraction problem. Our inter-NP causal relation extraction shows the precision of 81.29%, that is 7.05% improvement over the baseline model. The proposed models are also applied to inter-sentence causal relation extraction.

1 Introduction

Causality or *causal relation* refers to 'the relation between a cause and its effect or between regularly correlated events'[1]. Even if not a few questions order to find causality from text, the current Question Answering system cannot respond causal questions. The recent Question-Answering system can produce correct answers to 83.0% of questions (Moldovan et al., [2002]). But the answer accuracy has a wide variation across the question type. Moldovan et al. ([2003]) states the relatively high answer accuracy on questions about the person, the time, the location, and so on. They show very low performance on causal questions. Causal questions are answered with a low precision score of 3.1%. Since there are few causal questions in their test suite of TREC(Text REtrieval Conference), total performance is high in spite of the low performance on causal questions. However, causal questions are very frequently used in an actual question answering. For a web site[2] in which users exchanged questions and answers, there are

* This research was supported in part by KISTEP Strategic National R&D Program under contract M1-0107-00-0018 and by KOSEF under contract R21-2003-000-10042-0.

[1] From Merriam-Webster's Online Dictionary.

[2] Naver Knowledge iN, http://kin.naver.com

K.-Y. Su et al. (Eds.): IJCNLP 2004, LNAI 3248, pp. 61–70, 2005.

130,000 causal questions from 950,000 sentence-sized DB. This fact shows that it is necessary to analyze the causal relation for a high performance Question-Answering.

To response the causal question such as (1a), the following problems should be solved. The first problem is *event extraction*, which extracts events from the paragraphs including keyword 'hiccups' of the question. *Event* is defined as 'a phenomenon or occurrence located at a single point in space-time'[3]. The second problem is *causal relation extraction*, which analyzes the causal relation between events. *Causal question answering* is the last one to be solved. It infers the answer to the question. In this paper, we concentrate the causal relation extraction.

(1a) What are hiccups caused by?
(1b) The oral bacteria that cause gum disease appear to be the culprit.

Cue phrase is a word, a phrase, or a word pattern, which connects one event to the other with some relation. The causal relation between events is assumed by the cue phrase. The causal cue phrase is used for connecting the cause and effect events. When events are expressed by noun phrases, the cue phrase connecting events is a verb phrase in general. For example, in (1b), the verb 'cause' is a cue phrase to connect two events expressed by noun phrases, 'the oral bacteria' and 'gum disease'. Several lexical pairs are assumed to lead the causal relation. The lexical pair 'bacteria' and 'disease' is an example of the causal lexical pair. If the term pair 'the oral bacteria' and 'gum disease' is causally related, we can infer that the event pair 'bowel bacteria' and 'bowel disease' is causally related. Causal lexical pairs are learned from cause-effect event pairs. We define *lexical pair probability* as the probability of the lexical pair that is a part of causal event pairs. The pair of concept classes of each event, [B03] and [C23.550.288][4], also lead the causality of event pair. *Conceptual pair probability* is defined as the probability of the conceptual pair that has the causal relation. Cue phrases connecting two events are also considered to have connection probability. We define *cue phrase probability* as the probability of the cue phrase that connects causal event pairs. With these probabilities, we introduce a causal relation classifier based on Naive Bayes classifier. These probabilities are learned from the raw corpus in an unsupervised manner.

In section 2, selected works are compared for the causal relation extraction. Our classification model will be explained in section 3. In this paper, we aim to extract the causal relation that exists between two noun phrases or sentences. In section 4, we evaluate the proposed model for the inter-noun phrase causality extraction, and prove it adaptable to the inter-sentence causality extraction.

[3] From The American Heritage Dictionary of the English Language: Fourth Edition, 2000.

[4] They represent [bacteria] and [disease]. These conceptual numbers follow the biomedical ontology (Medical Subject Heading, [2004]).

2 Related Works

Causal relations are expressed with various forms in the literature: between subject and object noun phrases like (2a), between two sentences or phrases as in (2b,c), and in intra-structure of a noun phrase like (2d)[5]. The causal relation also exists between paragraphs that describe events. This relation is a part of rhetorical structure and is out of the focus in this paper. In the examples, each cause event is connected with its effect event by the causative verb ('generate'), causal connectives ('for this reason' or 'that'), or the intra-NP structure. In this paper, we are focusing on the inter-NP and inter-sentence causal relation.

(2a) Earthquakes generate tidal waves.
(2b) The meaning of a word can vary a great deal depending on the context. For this reason, pocket dictionaries have a very limited use.
(2c) The traffic was so heavy that I couldn't arrive on time.
(2d) Disease-causing bacteria

Marcu and Echihabi ([2002]) used the inter-sentence lexical pair probability for discriminating the rhetorical relation between sentences. To distinguish the causal relation from the other rhetorical relations, they used the sentence pairs connected with 'Because of' and 'Thus'. From the selected sentence pairs, causal lexical pairs are automatically collected. They used nouns, verbs, and adverbs only. Non-causal lexical pairs are also collected from randomly selected sentence pairs. With these two kinds of lexical pairs, they compose the Naive Bayes classifier. The sentence pairs are classified to 'causal' or not. The result showed 57% accuracy in the inter-sentence causality extraction. The lexical probability contributes to the causality extraction in part. From their result, we supposed that the lexical pair and other probabilities contribute to the inter-noun phrase causality extraction and the inter-sentence causality extraction. The other supposed probabilities are the cue phrase probability, cue phrase confidence score, and concept pair probability. With these probabilities and the unsupervised learning technique, we try to resolve the causality extraction problem.

Initial works on the causal relation analysis used hand-made causal patterns to find the causality. Khoo et al. ([1988] and [2000]) used the semi-automatic causality pattern learning on the syntactically analyzed corpus. Girju and Moldovan ([2002]) used the inter-noun phrase causal relation to improve the Question-Answering performance. To extract the inter-noun phrase causal relation, they used the cue phrase filter and the dictionary based ranking model. With simple rules reflecting the meaning of head nouns, each noun phrase pair is classified into 5 ordered classes. We call this order the *noun class rank*. The examination regarding classes 1 to 4 as causal relation shows a precision score of 65.5%. Their decision tree classifier learned on the causality-annotated corpus showed a precision of 73.91% (Girju, [2003]). In their works, cue phrases were verbs connecting a subject and an object noun phrase. 60 cue phrases were semi-automatically acquired from WordNet (Miller, [1995]) and corpus.

[5] Examples are selected from (Girju and Moldovan, [2002]).

For the supervised learning of causal relation classifier, causality-annotated corpus is required. But the construction of such corpus would take much effort. The supervised method has the limitation to be scaled up. In this paper, we propose new methods using the cue phrase probability and the lexical pair probability. The probabilities are learned from the raw corpus in an unsupervised manner. Proposed models use a large amount of corpus for improving the performance. In case of using the dictionary or WordNet as a basis of causality, the unregistered words in the dictionaries hinders from finding the correct causal relation. In manually annotated test set of section 4, 36.4% of unknown words decrease the performance of baseline systems. We solve this unknown word problem by using the lexical pair probability.

3 Causal Relation Extraction Model

3.1 Causality Candidate Extraction from Dependency Structure

In this works, the result of the event extraction is candidates of the event. After causal relation extraction, each event candidate is assigned to be a cause event, an effect event, or none of both. In the case of inter-noun phrase causality extraction, the event extraction is the noun phrase-chunking problem. For extracting event candidate, the syntactic analyzer is used. As an input of classifier, we use the ternary expression composed of cause-effect event candidates and cue phrases. For inter-noun phrase causal relation extraction, the ternary expression for causality candidates is <cause noun phrase, cue phrase, effect noun phrase>. Causality candidates are extracted from the dependency structure of sentences. We use several modules, which are a noun phrase chunking, a noun reference resolution, an appositive noun phrase analysis, and a cue phrase-filtering module.

Figure 1 is a dependency structure of the sentence (3a). It is the result of the Connexor dependency parser (Tapanainen and Jarvinen, [1997]). Noun phrases and transitive verbs are selected from the dependency tree. Intransitive verbs with a prepositional phrase are also considered. (Some prepositions introducing time and place are not considered.) The relative pronoun is replaced with its antecedent. We try to find the forward references in the boundary of a sentence tree. After the appositive noun phrase analysis is finished, we can find two ternary expressions, (3b,c), from the sentence (3a). These ternary expressions are finally filtered by pre-defined cue phrases. Pre-defined cue phrases are based on 60 causal verbs defined in (Girju et al., [2002]).

(3a) Skin cancer usually appears in adulthood, but it is caused by sun exposure and sunburns that began in childhood.
(3b) <'sun exposure', 'caused by', 'skin cancer'>
(3c) <'sunburn', 'caused by', 'skin cancer>

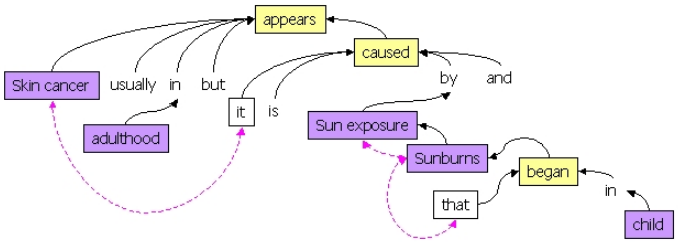

Fig. 1. Dependency structure of sentence (3a).

3.2 Causal Relation Classifier

The causality candidate, the ternary (t_i), is classified to be 'causal (c_1) ' or 'non-causal (c_0)'. To solve this classification problem, we apply Naive Bayes classifier. The class c^* of the ternary t_i is computed as shown in (1).

$$c^* = \arg\max_{c_j} P(c_j|t_i) = \arg\max_{c_j} \frac{P(c_j)P(t_i|c_j)}{P(t_i)} \tag{1}$$

When we consider the cue phrase CP_{t_i} and lexical pairs LP_{t_i}, as causal features of the ternary, $P(t_i|c_j)$ in (1) will be rewritten as (2). We assume these features are independent each other.

$$P(t_i|c_j) = P(CP_{t_i}|c_j) \prod_{k=1}^{|t_i|} P(LP_{t_i k}|c_j) \tag{2}$$

In (2), $P(CP_{t_i}|c_j)$ and $P(LP_{t_i k}|c_j)$ is the cue phrase probability and the lexical pair probability, which are defined in section 1. These probabilities can be learned from the causality-annotated ternary set. But the construction of the causality-annotated ternary set takes time and effort consuming. In this paper, we use the raw corpus rather than the causal relation annotated corpus. To make it possible, EM (Expectation-Maximization) procedure is used with the Naive Bayes classifier. Parameters trained in EM are prior probability $P(c_j)$, cue phrase probability $P(CP_{t_i}|c_j)$, and lexical pair probability $P(LP_{t_i k}|c_j)$. The parameters are smoothed using the Laplace method for the lexical pairs unseen in training data.

The Naive Bayes classifier is bootstrapped from the initial classifier. The training data is the ternary set filtered by cue phrases. There are three training stages. In the initialization stage, we build an initial classifier and initialize Naive Bayes parameters. As an initial classifier, we use the dictionary-based classifier described in section 2.3. It does not need the extra training sequence. The cue phrase confidence score is another available initial classifier. The cue phrase confidence score, $P(c_j|CP_{t_i})$, is defined as the probability of the causal class for the given cue phrase. This confidence score requires relatively small set of annotated corpus rather than the lexical pair probability. After whole training

corpus is classified with an initial classifier, highly ranked ternaries are selected as the initial causality-annotated set. From this annotated set, the parameters of Naive Bayes classifier are initialized.

The second training stage is called Expectation step. Whole training corpus, including the annotated part, is classified with the current classifier. The final training stage is called Maximization step. From the newly classified data, parameters are re-estimated. Expectation and Maximization step are repeated while classifier parameters improve.

3.3 Causality Classification Model

The trained classifier can be interleaved with other causal classifiers, which are the dictionary based rank (noun class rank) and the cue phrase confidence score. We propose three classification models. *The classification model CP+LP* uses the cue phrase probability and the lexical pair probability as shown in (3). This uses the noun class rank as a back-off model. If the cue phrase probability and the lexical pair probability, $P(c_j|t_i)$, cannot decide an evident causal class, the noun class rank probability is used. To do this, Discrimination value, $Dist(t_i)$, is introduced as shown in (4). The threshold h is a constant. The noun class rank probability $P(c_j|rank_{t_i})$ is defined as the probability of the causal class for the noun class rank of the given ternary. This probability is learned from the automatically annotated corpus.

$$P_{CP+LP}(c_j|t_i) = \begin{cases} P(c_j|t_i) & \text{if } Dist(t_i) > h \\ P(c_j|rank_{t_i}) & \text{otherwise} \end{cases} \tag{3}$$

$$Dist_{t_i} = \left| \frac{\log P(c_0|t_i) - \log P(c_1|t_i)}{\log P(c_0|t_i) + \log P(c_1|t_i)} \right| \tag{4}$$

The classification model CP+NC+LP uses the cue phrase probability and the lexical pair probability interleaved with the noun class rank probability as shown in (5). The sum of weights, w_{nc} and w_{lp}, must be 1.

$$P_{CP+NC+LP}(c_j|t_i) = w_{nc} \times P(c_j|rank_{t_i}) + w_{lp} \times P(c_j|t_i) \tag{5}$$

The cue phrase confidence score is also learned on the automatically annotated corpus. *The classification model CP+NC+CPC+LP* uses the cue phrase probability and the lexical pair probability interleaved with the noun class rank and the cue phrase confidence score as shown in (6). The sum of weights, w_{nc}, w_{cpc}, and w_{lp}, must be 1.

$$P_{CP+NC+CPC+LP}(c_j|t_i) = w_{nc} \times P(c_j|rank_{t_i}) + w_{cpc} \times P(c_j|CP_{t_i}) + w_{lp} \times P(c_j|t_i) \tag{6}$$

Figure 2 is the structure of proposed causality extraction system. Candidates for the causal event pairs are extracted from the raw corpus through an NP chunking with syntactic analysis. The causality candidate is composed of a pair of events and a cue phrase connecting two events. The causality analysis problem

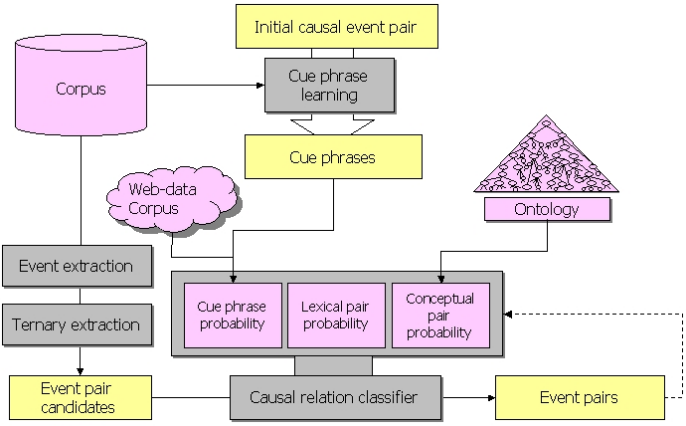

Fig. 2. Proposed causality extraction system.

is redefined as a classification problem to assign the causal class, 'causal' or 'non-causal', to the causality candidates. Causal event pairs are extracted with the cue phrase and lexical pair probabilities. Extracted causal event pairs are used for re-training the classifier.

4 Evaluation

4.1 Inter-NP Causality Extraction

A part of TREC corpus is used for the inter-noun phrase causality extraction. The training corpus is 5 million sentence-sized articles from LA TIMES (1989∼1990) and Wall Street Journal (1987∼1990). We use two test sets, which are selected from different domains. The first one is from Wall Street Journal articles. The other is from Medline medical encyclopedia of A.D.A.M. Inc. All sentences in test sets include the word 'cancer'. The first one, we call it cTREC, comes from general domain. The other, we call it cADAM, comes from medical domain. Test sets are manually classified with two human annotators, which one is the first author and the other is medical domain expert. They agree the result with 72.8%. A gold standard is made with discussion between annotators.

The cue phrase probability (CP) and the lexical pair probability (LP) are trained on the training set. As an initial classifier for the inter-noun phrase causality extraction, the noun class rank was used. For the parameter initialization, all ternaries were ranked with noun classes. And highly ranked ternaries were selected as a causality-annotated set. As a result, ternaries ranked by 1 to 3 were annotated to 'causal' (c_1), and parts of ternaries ranked by 5 were annotated to 'non-causal' (c_0). Table 1 shows the evaluation result on test sets. The classification model NC follows the model of (Girju and Moldovan, 2002), which uses the cue phrase filter and the noun class rank. The classification model

Table 1. Inter-noun phrase causal relation extraction result.

Classification model	Test set	Precision	Recall	F-value
NC	cTREC	82.88	64.79	72.73
	cADAM	65.17	42.34	51.33
	Total	75.00	53.76	62.63
LP with No EM	cTREC	76.35	79.58	77.93
	cADAM	73.68	51.09	60.34
	Total	75.71	67.03	71.10
CP+LP	cTREC	82.14	80.99	81.56
	cADAM	78.99	79.56	79.27
	Total	80.58	80.29	80.43
CP+NC+LP	cTREC	83.10	83.10	83.10
	cADAM	77.78	76.64	77.21
	Total	80.51	79.93	80.22
CP+NC+CPC+LP	cTREC	83.21	80.28	81.72
	cADAM	79.43	81.75	80.58
	Total	81.29	81.00	81.15

LP with No EM follows the classification model of (Marcu and Echihabi, 2002), which uses the lexical pair probability without EM process. The last three models are proposed models. For the classification model CP+LP, we assign 0 to the value of the threshold h. And for the noun class (NC) weight w_{nc} and the cue phrase confidence score (CPC) weight w_{cpc}, 0.1 is assigned.

Contribution of the cue phrase probability and the lexical pair probability: The proposed model CP+NC+CPC+LP shows the highest precision of 81.29%, which is improved by 7.05% from the baseline model (NC). Actually in all the proposed models, the causality extraction performance is increased. We can say that the cue phrase probability and the lexical pair probability are useful for the causality extraction.

Contribution of the noun class on domains: For the general domain test set (cTREC), the result of the interleaved with the noun class (CP+NC+LP) shows the precision improved by 1.19% from the non-interleaved (CP+LP). However, for the medical domain test set (cADAM), the precision is decreased by 1.53%. It is caused by unknown words in the medical domain test set. Terminologies and pronouns in the specific domain include more unknown words than in the general domain. For the baseline model NC, unknown words of cADAM decreases the performance by 15.1% in the precision and by 11.1% in the recall. We can say that the noun class is useful in general but not in the specific domain.

Contribution of the cue phrase confidence score: The classification model interleaved with the cue phrase confidence score (CP+NC+CPC+ LP) does not show the significant improvement from the non-interleaved model (CP+ NC+LP). It is because the cue phrase probability and the cue phrase confidence score share the same information space.

Robustness of the proposed model: In the proposed model (CP+LP), 37.5% of the unknown word-causing error of the baseline system (NC) is correctly classified. The proposed model does not refer the word sense. It refers only the lexical pair frequency in the corpus. We can say that the proposed model is free from the unknown words.

High performance of the unsupervised learning: The proposed models are learned in an unsupervised manner. It does not require the pre-annotated data. Nevertheless, the performance is relatively high.

4.2 Inter-sentence Causality Extraction in Korean

If events are represented by sentences or verb phrases leading subordinate clause, cue phrase could be conjunctive adverbs or verb endings in Korean. 54 cue phrase patterns are selected by human annotators. The event extraction module marks event boundaries on the dependency structure of sentences. In the inter-sentence causality extraction, an event is 'a predicate and its arguments'. After all event candidates are extracted from the dependency tree of adjacent sentence pairs, they are filtered by cue phrases. The input of the classifier is the ternary expression of <cause sentence, cue phrase, effect sentence>. We use the Korean dependency parser of (Chang and Choi, [2000]). The same classifier described in section 3 is used for the inter-sentence causality extraction. We use the cue phrase confidence score as the initial classifier. To learn this initial classifier, human annotator annotates 970 sentence pairs, which are 5∼20 sentence pairs for each cue phrase. The training set is the 2158 document sized medical encyclopedia (HealthChosun [2003], Joins HealthCare [2003]). It contains 30 thousand of ternaries. As a test set, we use 4 documents, which are not included in training data but same domain.

The universality of the proposed model: For the inter-sentence causality extraction, the proposed model (CP+LP) shows a precision of 74.67%, which is increased by 8.87% from the baseline (CPC). As a result, we can say that the proposed model is adaptable for the inter-sentence causal relation extraction.

5 Conclusion

In this paper, causality extraction models are proposed. Proposed models use the cue phrase probability and the lexical pair probability. These probabilities are learned from the raw corpus in unsupervised manner. Proposed models show higher performance than baseline systems. Proposed models use only the raw corpus and do not make a performance decrease by the unknown words. Therefore, the models can be easily adapted to the other domain such as the medical. Proposed models are used not only for the inter-noun phrase causality extraction but also for the inter-sentence causality extraction.

To use the conceptual pair probability, the word sense disambiguation has to be also considered. Proposed models classified the ternaries after filtered by

cue phrases. For the fully automatic learning of causality extraction, the cue phrase learning has to be solved. The incremental learning of the cue phrase is in progress. The proposed causality extraction is used for the causal Question Answering. The Question Answering based on causal relation is also going on. The causality extraction can be used for the causal browsing. The causal QA and browser is demonstrated on the web site, http://gensum.kaist.ac.kr/~dschang/ENC/CQA.html

References

[2000] Chang, Du-Seong and Key-Sun Choi, 2000, Unsupervised learning of the dependency grammar using inside and outside probabilities, in *Proceedings of the 12th Hangul and Korean Information Processing* (in Korean)

[2003] Girju, Roxana, 2003, Automatic Detection of Causal Relation for Question Answering, in *Proceeding of Workshop in the 41st Annual Meeting of the Association for Computational Linguistics Conference*

[2002] Girju, Roxana and Dan Moldovan, 2002, Mining Answers for Causation Questions, in *Proceeding of AAAI Symposium on Mining Answers from Texts and Knowledge Bases*

[2003] HealthChosun Medical Library, http://hpsearch.drline.net/dizzo/healthinfo/healthinfo.asp

[2003] Joins HealthCare Medical Encyclopedia, http://healthcare.joins.com/library

[2000] Khoo, Cristopher S.G., Syin Chan, and Yun Niu, 2000, Extracting Causal Knowledge from a Medical Database Using Graphical Patterns, In *Proceedings of The 38th Annual Meeting of the Association for Computational Linguistics*

[1988] Khoo, Cristopher S.G., J. Kornfit, Robert N. Oddy, and Sung Hyon Myaeng, 1998, Automatic Extraction of Cause-Effect Information from Newspaper Text without Knowledge-Based Inferencing, in *Literary and Linguistic Computing*, 13(4), pages 177-186

[2002] Marcu, Daniel and Abdessamad Echihabi, 2002, An Unsupervised Approach to Recognizing Discourse Relations, in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics Conference*, Philadelphia, PA

[2004] Medical Subject Heading, 2004, http://www.nlm.nih.gov/mesh

[1995] Miller, G., 1995, WordNet: a Lexical Database, *Communications of the ACM*, 38(11):39-41

[2003] Modovan, Dan I., Marius Pasca, Sanda M. Harabagiu, Mihai. Surdeanu, 2003, Performance Issues and Error Analysis in an Open-Domain Question Answering, *ACM Transactions on Information Systems*, Vol. 21, No. 2, pages 133-154.

[2002] Moldovan, Dan I., Sanda M. Harabagiu, Roxana Girju, Paul Morarescu, Finley Lacatusu, Adrian Novischi, Adriana Badulescu and Orest Bolohan, 2002, LCC Tools for Question Answering, in *Proceedings of the 11th Text Retrieval Conference*, NIST.

[2000] Nigram, Kamal, Andrew K. McCallum, Sebastian Thrun and Tom Mitchell, 2000, Text Classification from Labeled and Unlabeled Documents using EM, *Machine Learning*, 39(2/3) pages 103-134.

[1997] Tapanainen, Pasi and Timo Jarvinen, 1997, A non-projective dependency parser in Proceedings of the 5th Conference on Applied Natural Language Processing, *Association for Computational Linguistics*, pages 64-71.