Efficient Dependency-Guided Named Entity Recognition

Zhanming Jie, Aldrian Obaja Muis, Wei Lu
Singapore University of Technology and Design
8 Somapah Road, Singapore, 487372
zhanming.jie@mymail.sutd.edu.sg, {aldrian_muis,luwei}@sutd.edu.sg

Abstract

Named entity recognition (NER), which focuses on the extraction of semantically meaningful named entities and their semantic classes from text, serves as an indispensable component for several downstream natural language processing (NLP) tasks such as relation extraction and event extraction. Dependency trees, on the other hand, also convey crucial semantic-level information. It has been shown previously that such information can be used to improve the performance of NER (Sasano and Kurohashi 2008; Ling and Weld 2012). In this work, we investigate on how to better utilize the structured information conveyed by dependency trees to improve the performance of NER. Specifically, unlike existing approaches which only exploit dependency information for designing local features, we show that certain global structured information of the dependency trees can be exploited when building NER models where such information can provide guided learning and inference. Through extensive experiments, we show that our proposed novel dependency-guided NER model performs competitively with models based on conventional semi-Markov conditional random fields, while requiring significantly less running time.

Introduction

Named entity recognition (NER) is one of the most important tasks in the field of natural language processing (NLP). The task focuses on the extraction of named entities together with their semantic classes (such as organization or person) from text. The extracted named entity information has been shown to be useful in various NLP tasks, including coreference resolution, question answering and relation extraction (Lao and Cohen 2010; Lee et al. 2012; Krishnamurthy and Mitchell 2015).

Dependency trees, on the other hand, were shown to be useful in several semantic processing tasks in NLP such as semantic parsing and question answering (Poon and Domingos 2009; Liang, Jordan, and Klein 2013). The dependency structures convey semantic-level information which was shown to be useful for the NER task. Existing research efforts have exploited such dependency structured information by designing dependency-related local features that can be used in the NER models (Sasano and Kurohashi 2008; Ling and Weld 2012; Cucchiarelli and Velardi 2001). Figure 1 shows two example phrases annotated with both dependency and named entity information. The edges on top of words represent the dependencies and the labels with IOB encoding are the entity types.
information conveyed by dependency trees for building novel models for improved named entity recognition. The model assumes the availability of dependency trees before performing NER, which can be obtained from a dependency parser or given as part of the input. Unlike existing approaches that only exploit dependency structures for encoding local features, the model is able to explicitly take into account the global structured information conveyed by dependency trees when performing learning and inference. We call our proposed NER model the dependency-guided model (DGM), and build it based on the conventional semi-Markov conditional random fields (semi-CRFs) (Sarawagi and Cohen 2004), a classic model used for information extraction.

Our main contributions can be summarized as follows:

- We present a novel model that is able to explicitly exploit the global structured information conveyed by dependency trees, showing that such information can be effectively integrated into the process of performing NER. To the best of our knowledge, this is the first work that exploits such information for NER.
- Theoretically, we show through average-case time complexity analysis that our model has the same time complexity as that of the linear-chain CRFs, and is better than that of the semi-Markov CRFs.
- Empirically, we demonstrate the benefits of our approach through extensive experiments on benchmark datasets. We show that the resulting model leads to NER results that are competitive with the baseline approach based on semi-Markov CRFs, while requiring significantly less running time.

Furthermore, although in this paper we focus on the task of using the structured information conveyed by dependency trees for improved NER, the underlying idea is general, which we believe can be applied to other tasks that involve building pipeline or joint models for structured prediction. Potentially if we can find the relationship between different NLP tasks, we can build an efficient model for a specific task while using information from other tasks without making the model more complex.

Related Work

Named entity recognition has a long history in the field of natural language processing. One standard approach to NER is to regard the problem as a sequence labeling problem, where each word is assigned a tag, indicating whether the word belongs to part of any named entity or appears outside of all entities. Previous approaches used sequence labeling models such as hidden Markov models (HMMs) (Zhou and Su 2002), maximum entropy Markov models (MEMMs) (McCallum, Freitag, and Pereira 2000), as well as linear-chain (Finkel, Grenager, and Manning 2005) and semi-Markov conditional random fields (CRFs/semi-CRFs) (Sarawagi and Cohen 2004). Muis and Lu (2016b) proposed a weak semi-CRFs model which has a lower complexity than the conventional semi-CRFs model while still having a higher complexity than the linear-chain CRFs model. Our model is proved to have the same time complexity as linear-chain CRFs model in the average case. The quality of the CRFs model typically depends on the features that are used. While most research efforts exploited standard word-level features (Ratinov and Roth 2009), more sophisticated features can also be used. Ling and Weld (2012) showed that using syntactic-level features from dependency structures in a CRFs-based model can lead to improved NER performance. Such dependency structures were also used in the work by Liu, Huang, and Zhu (2010), where the authors utilized such structures for building a skip-chain variant of the original CRFs model. This shows that some simple structured information conveyed by dependency trees can be exploited for improved NER. In their skip-chain CRFs model, they simply added certain dependency arcs as additional dependencies in the graphical model, resulting in loopy structures. However, such a model did not explicitly explore the relation between entities and global structured information of the dependency trees. The authors also showed that such a model does not outperform a simpler approach that adds additional dependencies between similar words only on top of the original CRFs model. In this work, we also focus on utilizing dependency structures for improving NER. Unlike previous approaches, we focus on exploiting the global structured information conveyed by dependency trees to improve the NER process. Comparing with the semi-CRFs model, our model is not only able to perform competitively in terms of performance, but also more efficient in terms of running time.

There are also some existing works that focus on improving the efficiency of NER and other information extraction models. For example, Okanohara et al. (2006) used a separate naive Bayes classifier to filter some entities during training and inference in their semi-CRFs based model. While the filtering process was used to reduce the computational cost of the semi-CRFs model, the model still needs to enumerate all the possible chunks. Yang and Cardie (2012) extended the original semi-CRFs for extracting opinion expressions and used the constituency parse tree information to avoid constructing implausible segments. Lu and Roth (2015) proposed an efficient and scalable model using hypergraph which can handle overlapping entities. Muis and Lu (2016a) extended the hypergraph representation to recognize both contiguous and discontiguous entities.

Background

Before we introduce our models, we would like to have a brief discussion on the relevant background. Specifically, in this section, we review the two classic models that are commonly used for named entity recognition, namely the linear-chain conditional random fields and the semi-Markov conditional random fields models.

Linear-chain CRFs

Conditional random fields, or CRFs (Lafferty, McCallum, and Pereira 2001) is a popular model for structured prediction, which has been widely used in various natural language processing problems, including named entity recognition (McCallum and Li 2003) and semantic role labeling (Cohn and Blunsom 2005).
We focus our discussions on the linear-chain CRFs in this section. The probability of predicting a possible output sequence $y$ (e.g., a named entity label sequence in our case) given an input $x$ (e.g., a sentence) is defined as:

$$p(y|x) = \frac{\exp(w^Tf(x, y))}{Z(x)}$$  (1)

where $f(x, y)$ is the feature vector defined over $(x, y)$ tuple, $w$ is the weight vector consisting of parameters used for the model, and $Z(x)$ is the partition function used for normalization, which is defined as follows:

$$Z(x) = \sum_{y'} \exp(w^Tf(x, y'))$$  (2)

In a (first-order) linear-chain CRF, the partition function for a input of length $n$ can also be written as follows:

$$Z(x) = \sum_{i=1}^{n} \sum_{y' \in T} \exp(w^Tf(x, y', y, i))$$  (3)

where $f(x, y', y, i)$ is a local feature vector defined at the $i$-th position of the input sequence. $T$ is the set of all output labels. The above function can be computed efficiently using dynamic programming. It can be seen that the time complexity of a linear-chain CRFs model is $O(n|T|^2)$.

We aim to minimize the negative joint log-likelihood with $L_2$ regularization for our dataset:

$$\mathcal{L}(w) = \sum_{i} \log \sum_{y'} \exp(w^Tf(x_i, y'))$$

$$- \sum_{i} w^Tf(x_i, y_i) + \lambda w^T w$$  (4)

where $(x_i, y_i)$ is the $i$-th training instance and $\lambda$ is the $L_2$ regularization coefficient.

The objective function is convex and we can make use of the L-BFGS (Byrd et al. 1995) algorithm to optimize it. The partial derivative of $\mathcal{L}$ with respect to each parameter $w_k$ is:

$$\frac{\partial \mathcal{L}}{\partial w_k} = \sum_{i} \left( E_{p(y'|x_i)}[f_k(x_i, y')] - f_k(x_i, y_i) \right) + 2\lambda w_k$$

Semi-Markov CRFs

The semi-Markov conditional random fields, or semi-CRFs (Sarawagi and Cohen 2004) is an extension of the standard linear-chain CRFs. Different from linear-chain CRFs, which makes use of simple first-order Markovian assumptions, the semi-CRFs assumes that the transitions between words inside a span (e.g., an entity consisting of multiple words) can be non-Markovian. Such an assumption allows more flexible non-Markovian features to be defined. The resulting model was shown to be more effective than its linear-chain counterpart in information extraction tasks.

The partition function in this setting becomes:

$$Z(x) = \sum_{i=1}^{n} \sum_{l=1}^{L} \sum_{y' \in T} \sum_{y \in T} \exp(w^Tf(x, y', y, i - l, i))$$  (5)

where $f(x, y', y, i - l, i)$ represents the local feature vector at position $i$ with a span of size $l$. In this case, the time complexity becomes $O(nL|T|^2)$. The value $L$ is the maximal length of the spans the model can handle.

Consider the sentence “Lee Ann Womack won Single of the Year award”. The upper portion of Figure 2 shows all the possible structures that are considered by the partition function with $L = 4$. These structures essentially form a compact lattice representation showing all the possible combinations of entities (with length restriction) that can ever appear in the given sentence. Each orange and red edge is used to represent an entity of type PER and MISC respectively. The black lines are used to indicate words that are not part of any entity. The directed path highlighted in bold blue corresponds to the correct entity information associated with the phrase, where “Lee Ann Womack” is an entity of type PER, and “Single of the Year” is another entity of type MISC.

Our Model

The primary assumption we would like to make is that the dependency trees of sentences are available when performing the named entity recognition task.

NER with Dependency Features

One approach to exploit information from dependency structures is to design local features based on such dependency structures and make use of conventional models such as semi-CRFs for performing NER. Previous research efforts have shown that such features extracted from dependency structures can be helpful when performing the entity recognition task (Sasano and Kurohashi 2008; Ling and Weld 2012; Cucchiarelli and Velardi 2001).

In Figure 2, we have already used a graphical illustration to show the possible combinations of entities that can ever appear in the given sentence. Each edge in the figure corresponds to one entity (or a single word that is outside of any entity – labeled with $o$). Features can be defined over such edges, where each feature is assigned a weight. Such features, together with their weights, can be used to score each possible path in the figure. When dependency trees are provided, one can define features involving some local dependency structures. For example, for the word “Womack”, one possible feature that can be defined over edges covering this word can be of the form “Womack ← won”, indicating there is an incoming arc from the word “won” to the current word “Womack”. However, we note that such features largely ignore the global structured information as presented by the dependency trees. Specifically, certain useful facts such as the word “Lee” is a child of the word “Womack” and at the same time a grandchild of the word “won” is not captured due to such a simple treatment of dependency structures.

Dependency-Guided NER

The main question we would like to ask is: how should we make good use of the structured information associated with the dependency trees to perform named entity recognition? Since we are essentially interested in NER only, would there
be some more global structured information in the dependency trees that can be used to guide the NER process?

From the earlier two examples shown in Figure 1 as well as the example shown in Figure 2 we can observe that the named entities tend to be covered by a single or multiple adjacent dependency arcs. This is not a surprising finding as for most named entities which convey certain semantically meaningful information, it is expected that there exist certain internal structures – i.e., dependencies amongst different words within the entities. Words inside each named entity typically do not have dependencies with words outside the entities, except for certain words such as head words which typically have incoming arcs from outside words.

This finding motivates us to exploit such global structured information as presented by dependency trees for performing NER. Following the above observation, we first introduce the following notion:

**Definition 1** A valid span either consists of a single word, or is a word sequence that is covered by a chain of (undirected) arcs where no arc is covered by another.

For example, in Figure 2, the word sequence “Lee Ann Womack won Single of the Year” is not a valid span, since the word sequence is not covered by a chain of arcs. However, the word sequence “Lee Ann Womack” is a valid span since there exists a chain of arcs from “Lee” to “Ann” to “Womack” to “won”. Similarly, the single word “Womack” is also a valid span given the above definition.

Based on the above definition, we can build a novel dependency-guided NER model based on the conventional semi-CRFs by restricting the space of all possible combinations of entities to those that strictly contain only valid spans. This leads to the following new partition function:

$$Z(x) = \sum_{(i,j) \in S_L(x)} \sum_{y' \in T} \sum_{y \in T} \exp(w^T f(x, y', y, i, j))$$

where $S(x)$ refers to the set of valid spans for a given input sentence $x$, and $S_L(x)$ refers to its subset that contains only those valid spans whose lengths are no longer than $L$.

We call the resulting model dependency-guided model (DGM). Figure 2 (middle) presents the graphical illustration of all possible paths (combination of entities) with $L = 4$ that our model considers. For example, since the word sequence “Lee Ann Womack won Single” is not a valid span, the corresponding edges covering this word sequence is removed from the original lattice representation in Figure 2 (top). The edge covering the word sequence “of the Year” remains there as it is covered by the arc from “of” to “Year”, thus a valid span. As we can see, the amount of paths that we consider in the new model is substantially reduced.

**Time Complexity**

We can also analyze the time complexity required for this new model. We show example best-case and worst-case scenarios in Figure 3. For the former case there are $O(n)$ valid spans. Thus the time complexity in the best case is $O(nT^2)$, the same as the linear-chain CRFs model. For the latter, there are $O(nL)$ valid spans, leading to the time complexity $O(nL|T|^2)$ in the worst case, the same as the semi-Markov CRFs model.

Furthermore, we have the following theorem for the average-case time complexity:
Theorem 1 The average-case time complexity of DGM is $O(n|T|^2)$.

Proof We provide two proofs to the above theorem. Below we show one proof and we move the second proof to the supplementary material (S.3.1). In this proof we assume the largest possible value for $L$, which is $n$ (i.e., we do not limit the length of valid spans). Since the time complexity of our model depends directly on the total number of edges in the DGM graph, which is the same as the number of valid spans, we count the total number of valid spans over all possible structures by defining a bijection between each valid span and a tree with $(n + 1)$ nodes. Here we consider only one entity type since when we have $|T|$ types, the number of edges can be simply calculated by multiplying $|T|^2$ to the number that we will calculate here.

Let $\mathcal{T}(n)$ be the set of all undirected trees with $n$ nodes, let $\tau \in \mathcal{T}(n)$ be an undirected tree with $n$ nodes, let $s^u_v^{\tau}$ be a valid span based on $\tau$ covering the words from position $u$ to $v$, and let $S(\tau) = \{s^u_v^{\tau} \mid 1 \leq u \leq v \leq n\}$ be the set of valid spans based on $\tau$. Then the set of all valid spans over all possible undirected trees with $n$ nodes is $S_{DGM}(n) = \bigcup_{\tau \in \mathcal{T}(n)} S(\tau)$. Note that since the definition of valid spans uses undirected arcs of the dependency tree, we use undirected trees to count the number of valid spans.

Now we define the bijection from $S_{DGM}(n)$ to $\mathcal{T}(n+1)$ to show that the two sets have the same number of elements. By definition of valid spans, for each $s^u_v^{\tau} \in S_{DGM}(n)$, there is exactly one associated edges in the dependency tree $\tau$, which is the chain of edges starting at $u$ and ending at $v$, possibly visiting some intermediary nodes in between.

Then, to define the bijection we map $s^u_v^{\tau}$ to a tree $\tau' \in \mathcal{T}(n+1)$ which is constructed from $\tau$ as follows: the edges for the first $n$ nodes coincide with $\tau$, except that the edges involved in the chain of edges are removed, and we make the nodes involved in this chain of edges to be connected to the $(n+1)$-th node instead. The process creates a well-defined mapping since the resulting graph is still connected and has $n$ edges over $(n+1)$ nodes, which means it is a tree with $(n+1)$ nodes.

Conversely, from a tree $\tau' \in \mathcal{T}(n+1)$ we can recover the valid span and the undirected tree $\tau$ it is based on by removing the $(n+1)$-th node and the associated edges, then connecting the nodes that were previously connected to the $(n+1)$-th node in increasing order. The smallest and largest positions of the nodes define the starting and ending position of the span, and the resulting tree is the undirected tree $\tau$.

For illustration, Figure 4a shows an example of a valid span $s^1_4^{\tau}$ where $\tau \in \mathcal{T}(4)$ and the nodes and edges involved in defining the valid span are highlighted in red. Figure 4b shows the associated tree $\tau' \in \mathcal{T}(5)$ where the removed edges are put in dashed lines and the new edges in bold.

Therefore the mapping is bijective, and thus the number of total valid spans in $S_{DGM}(n)$ is equal to the number of trees in $\mathcal{T}(n + 1)$. It is a known result that the number of possible trees with $n$ nodes is $n^{n-2}$ (Aigner and Ziegler 2010, Chapter 30), which means the total number of valid spans is $n^{n-2}$.

Figure 4: An illustration of the bijection between a valid span $s^u_v^{\tau}$ and a tree $\tau' \in \mathcal{T}(n + 1)$.

\[
(n + 1)^{n-1} \leq n \cdot e
\]

which is linear in terms of $n$, just like the linear-chain CRFs. Here $e$ is the Euler number ($\approx 2.718$).

This shows that the average-case time complexity of our model is $O(n|T|^2)$.

Experimental Setup

For experiments, we followed (Finkel and Manning 2009) and used the Broadcast News section of the OntoNotes dataset. Instead of using its earlier 2.0 release, we used the final release – release 5.0 of the dataset, which is available for download\footnote{https://catalog.ldc.upenn.edu/LDC2013T19/}. The dataset includes the following 6 subsections: ABC, CNN, MNB, NBC, PRI and VOA. Moreover, the current OntoNotes 5.0 release also includes some English P2.5 data, which consists of corpus translated from Chinese and Arabic.\footnote{The OntoNotes 5.0 dataset also contains the earlier release from OntoNotes 1.0 to OntoNotes 4.0. However, we found the number of sentences of the Broadcast News in the current OntoNotes 5.0 dataset does not match the number reported in (Finkel and Manning 2009; 2010), which was based on OntoNotes 2.0. Furthermore, they removed some instances which are inconsistent with their model. We thus cannot conduct experiments to compare against the results reported in their work.} Following (Finkel and Manning 2009)
We provide more detailed statistics for each of the 7 subsections in the supplementary material (S.1). These numbers confirm that it is indeed the case that most named entities do form valid spans, even when using predicted dependency trees, and that such global structured information of dependency trees can be exploited for NER.

**Results and Discussions**

**NER Performance**

Following previous work (Finkel and Manning 2009), we report standard F-score in this section (detailed results with precision and recall can be found in the supplementary material S.5). Table 2 shows the results of all models when dependency features are exploited. Specifically, it shows results when the given and predicted dependency trees are considered, respectively. Overall, the semi-Markov CRFs, DGM-S and DGM models perform better than the linear-chain CRFs model. Our DGM model obtains an overall F-score at 80.5% and outperforms the semi-CRFs model significantly ($p < 0.001$ with bootstrap resampling (Koehn 2004)). For individual subsection, our DGM also performs the best. On 2 out of the 7 subsections, our model’s improvements over the baseline semi-CRFs model are significant ($p < 0.001$ with bootstrap resampling). For some other subsections like ABC, CNN, MNB, NBC and PRI, DGM has higher F-score than the semi-CRFs model but the improvements are not statistically significant. The results show that comparing with semi-CRFs, our DGM model, which has a lower average-case time complexity, still maintains a competitive performance. Such results confirm that the global structured information conveyed by dependency trees are useful and can be exploited by our DGM model.

The performance of DGM-S is worse than that of semi-CRFs and DGM in general since there are still many named entities that can not be handled by such a simplified model (see Table 1). This model has the same time complexity as the linear-chain CRFs, but performs better than linear-chain

---

1 http://stanfordnlp.github.io/CoreNLP/
2 http://maltparser.org/
3 https://catalog.ldc.upenn.edu/LDC2011T01/
CRFs, largely due to the fact that certain structured information of dependency trees are exploited in DGM-S. We note that such a simplified model obtains similar results as semi-CRFs for the two larger subsections, PRI and VOA. This is largely due to the fact that a larger percentage of the entities in these two subsections can be handled by DGM-S (see supplementary material S.1 for more details). It is noted that the performance of both semi-CRFs and DGM degrades when the predicted dependency trees are used instead of the given. The drop in F-score for DGM is more significant as compared to the semi-CRFs. Nevertheless, their results remain comparable. Such experiments show the importance of considering high quality dependency structures for performing guided NER in our model.

To understand the usefulness of the global structured information of dependency trees better, we conducted further experiments by excluding dependency features from our models. Such results are shown in Table 3. Our DGM model consistently performs competitively with the semi-CRFs model. The only exception occurs when the ABC subsection is considered and the predicted dependency trees are used ($p=0.067$). In general, we can see that when dependency features are excluded, the overall F-score for all models drop substantially. However, we note that for semi-CRFs, the drop in F-score is 1.1% for given dependency trees, and is 1.0% for predicted trees, whereas for DGM, the drops with given and predicted trees are both 0.6%. Overall, such results largely show that our proposed model is able to effectively make use of the global structured information conveyed by dependency trees for NER.

We have also conducted experiments on the widely-used NER dataset, CoNLL-2003, using the Stanford dependency parser\(^6\) to generate the dependency trees. Using the same feature set that we describe in this paper, our models do not achieve the state-of-the-art results on this dataset. However, they still perform comparably with the semi-CRFs model, while requiring much lesser running time. Note that since our goal in this paper is to investigate the usefulness of incorporating the dependency structure information for NER, we did not attempt to tune the feature set to get the best result on a specific dataset. Also it is worth remarking that we still obtain a relatively good performance for our DGM model although the dependency parser is not trained within the dataset itself and that a correct dependency structure information is crucial for the DGM model.

### Error Analysis

We provide a further analysis of how the dependency parsing performance affects NER based on Table 3. Because our model uses the dependency structure information directly in-

---

\(^6\)http://nlp.stanford.edu/software/nndep.shtml
stead of using them as features, we can analyze the influence of dependency structures on NER more clearly. Specifically, we focus on how the dependency parsing results affect the prediction of our DGM model.

A typical error made by our model taken from the results is shown in Figure 5 where the dependency tree in Figure 5a is given by the conversion of constituent parse tree and the other one in Figure 5b is predicted from the MaltParser. The NER result on the left is correct while the one on the right is incorrect. Based on the predicted dependency structure in Figure 5b, there is no way to predict an entity type for “USS Cole” since this is not a valid span in DGM model. Furthermore, DGM can actually recognize “Navy” as an ORG entity even though the predicted dependency is incorrect. But the model considers “USS Cole Navy” as an entity due to the interference of other entity features (e.g., NNP tag and Capitalized pattern) that “USS Cole” has. While with the given dependency tree, DGM considers “USS Cole” as a valid span and correctly recognizes it as a MISC entity.

**Speed Analysis**

From Figure 2 we can see that the running time required for each model depends on the number of edges that the model considers. We thus empirically calculated the average number of edges per word each model considers based on our training data. We found that the average number of edges involved in each token is 132 for the semi-CRFs model, while this number becomes 40 and 61 for DGM-S and DGM respectively. A lower number of edges indicates less possible structures to consider, and a reduced running time. See more detailed information in the supplementary material (S.3.2).

The results on training time per iteration (inference time) for all 7 subsections are shown in Figure 6. From the figure we can see that the linear-chain CRFs model empirically runs the fastest. The simple DGM-S model performs comparably with linear-chain CRFs. The semi-CRFs model requires substantially longer time due to the additional factor $L$ (set to 8 in our experiments) in its time complexity. In contrast, our model DGM requires only 47% of the time needed for semi-CRFs for each training iteration, and requires 37% more time than the DGM-S model.

**Conclusion**

In this paper, we proposed a novel efficient dependency-guided model for named entity recognition. Motivated by the fact that named entities are typically covered by dependency arcs, presenting internal structures, we built a model that is able to explicitly exploit global structured information conveyed by dependency trees for NER. We showed that the model theoretically is better than the semi-Markov CRFs model in terms of time complexity. Experiments show that our model performs competitively with the semi-Markov CRFs model, even though it requires substantially less running time. Future directions include further investigations on the structural relations between dependency trees and named entities, and working towards building integrated models that perform joint prediction of both structures.

We make our code and system available for download at [http://statnlp.org/research/ie/](http://statnlp.org/research/ie/).

**Acknowledgments**

We would like to thank the anonymous reviewers for their constructive and helpful comments. This work is supported by MOE Tier 1 grant SUTDT12015008.

**References**


Finkel, J. R.; Grenager, T.; and Manning, C. 2005. Incorporating non-local information into information extraction.