Transformation Based Chinese Entity Detection and Tracking

Yaqian Zhou*
Dept. Computer Science and Enginering
Fudan Univ.
Shanghai 200433, China
ZhouYaqian@fudane.edu.cn

Changning Huang
Microsoft Research, Asia
Beijing 100080, China
cnhuang@msrchina.research.microsoft.com

Jianfeng Gao
Microsoft Research, Asia
Beijing 100080, China
jfgao@microsoft.com

Lide Wu
Dept. of CSE., Fudan Univ.
Shanghai 200433, China
ldwu@fudane.edu.cn

Abstract

This paper proposes a unified Transformation Based Learning (TBL, Brill, 1995) framework for Chinese Entity Detection and Tracking (EDT). It consists of two sub models: a mention detection model and an entity tracking/coreference model. The first sub-model is used to adapt existing Chinese word segmentation and Named Entity (NE) recognition results to a specific EDT standard to find all the mentions. The second sub-model is used to find the coreference relation between the mentions. In addition, a feedback technique is proposed to further improve the performance of the system. We evaluated our methods on the Automatic Content Extraction (ACE, NIST, 2003) Chinese EDT corpus. Results show that it outperforms the baseline, and achieves comparable performance with the state-of-the-art methods.

1 Introduction

The task of Entity Detection and Tracking (EDT) is suggested by the Automatic Content Extraction (ACE) project (NIST, 2003). The goal is to detect all entities in a given text and track all mentions that refer to the same entity. The task is a fundamental to many Natural Language Processing (NLP) applications, such as information retrieval and extraction, text classification, summarization, question answering, and machine translation.

EDT is an extension of the task of coreference resolution in that in EDT we not only resolve the coreference between mentions but also detect the entities. Each of those entities may have one or more mentions. In the ACE project, there are five types of entities defined in EDT: person (PER), geography political Entity (GPE), organization (ORG), location (LOC), and facility (FAC). Many traditional coreference techniques can be extended to EDT for entity tracking.

Early work on pronoun anaphora resolution usually uses rule-based methods (e.g. Hobbs 1976; Ge et al., 1998; Mitkov, 1998), which try to mine the cues of the relation between the pronouns and its antecedents. Recent research (Soon et al., 2001; Yang et al., 2003; Ng and Cardie, 2002; Ittycherah et al., 2003; Luo et al., 2004) focuses on the use of statistical machine learning methods and tries to resolve references among all kinds of noun phases, including name, nominal and pronoun phrase. One common approach applied by them is to first train a binary statistical model to measure how likely a pair of

* This work is done while the first author is visiting Microsoft Research Asia.

mentions corefer; and then followed by a greedy procedure to group the mentions into entities.

Mention detection is to find all the named entity, noun or noun phrase, pronoun or pronoun phrase. Therefore, it needs Named Entity Recognition, but not only. Though the detection of entity mentions is an essential problem for EDT/coreference, there has been relatively less previous research. Ng and Cardie (2002) shows that improving the recall of noun phrase identification can improve the performance of a coreference system. Florian et al. (2004) formulate the mention detection problem as a character-based classification problem. They assign for each character in the text a label, indicating whether it is the start of a specific mention, inside a specific mention, or outside of any mention.

In this paper, we propose a unified EDT model based on the Transformation Based Learning (TBL, Brill, 1995) framework for Chinese. The model consists of two sub models: a mention detection model and a coreference model. The first sub-model is used to adapt existing Chinese word segmentation and Named Entity (NE) recognition system to a specific EDT standard. TBL is a widely used machine learning method, but it is the first time it is applied to coreference resolution. In addition, a feedback technique is proposed to further improve the performance of the system.

The rest of the paper is organized as follows. In section 2, we propose the unified TBL Chinese EDT model framework. We describe the four key techniques of our Chinese EDT, the word segmentation adaptation model, the mention detection model, the coreference model and the feedback technique in section 3, 4, 5 and 6 accordingly. The experimental results on the ACE Chinese EDT corpus are shown in section 7.

2 The Unified System Framework

Our Chinese EDT system consists of two components, mention detection module and coreference module besides a feedback technique between them as illustrated in Figure 1.

MSRSeg (Gao et al., 2003; Gao et al.), Microsoft Research Asia's Chinese word segmentation system that is integrated with named entity recognition, is used to segment Chinese words. However MSRSeg can't well match the standard of ACE EDT evaluation for either types or boundaries. The difference of the standard of named entity between MSRSeg and ACE cause more than half of the errors for NAME mention detection. In order to overcome these problems, we integrate a segmentation adapter to mention detection model.

The EDT system is a unified system that uses the TBL scheme. The idea of TBL is to learn a list of ordered rules while progressively improve upon the current state of the training set. An initial assignment is made based on simple statistics, and then rules are greedily learned to correct the mistakes, until no more improvement can be made. There are three main problems in the TBL framework: An initial state assignment, a set of allowable templates for rules, and an objective function for learning.

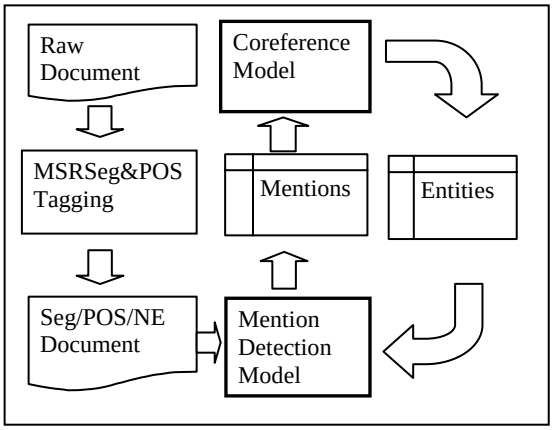

Figure 1. Entity detection and tracking system flow.

3 Word Segmentation Adaptation

The method of applying TBL to adapt the Chinese word segmentation standard has been described in Gao et al. (2004). Our approach is slightly different for not have a correctly segmented corpus according to ACE standard.

From the un-segmented ACE EDT corpus, we can only obtain mention boundary information. So the adapting objective is to detect the mention boundary instead of all words in text, correctly. In the corpus, very few mentions' boundaries are crossing[1].

The initial state of the segmentation adaptation model is the output of MSRSeg. And we

[1] The mentions' extents are frequently crossing, while heads not.

define two actions in the model, inserting and removing a boundary. The prefix or suffix of current word is used to define the boundary of inserting. Both inserting and removing action consider the combination of POS tag and word string of current, left and right words.

When inserting a boundary, the right part of the word keeps the old POS tag, and the left part introduces a special POS tag "new". When removing a boundary, the new formed word introduces a special POS tag "new". The following two examples illustrate the strategy.

俄 罗 斯 法 院 /nt/court of Russia → 俄 罗 斯 /**new**/Russia 法院/nt/court

波/nr/Bo 普/nr/Pu →波普/**new**/Bopu

4 Mention Detection

Since the word segmentation adaptation model has corrected the boundaries of mentions, our mention detection model bases on word and only tagging the entity mention types. The model detects the mentions by tagging sixteen tags (including the combination of five entity types and three mention types and "OTHER" tag) for all the words outputted by segmentation adaptation model. The templates, as illustrated in table 1, only refer to local features, such as POS tag and word string of left, right, and current words; the suffix, and single character feature of current word.

Table 1. Templates for mention detection.

MT1: P0	**MT9:** R4,P0
MT2: W0	**MT10:** R3,P0
MT3: P0,W0	**MT11:** R2,P0
MT4: P_1,W0	**MT12:** R1,P0
MT5: P_1,P0	**MT13:** S0,P0
MT6: W0,P1	**MT14:** T_1,W0
MT7: P0,P1	**MT15:** T_1,P0
MT8: W0,W1	**MT16:** P0,T1

Table 2. Examples of transformation rules of mention detection.

MR1: MT13 0 ns GPE
MR2: MT13 0 nr PER
MR3: MT13 0 nt ORG
MR4: MT16 n PER NPER
MR5: MT16 new ORG GPE

In table 1, "MT1" et al represent the id of the templates; "R1", "R2", "R3" and "R4" represent the suffix of current word and the number of character is 1, 2, 3 and 4 accordingly; other suffix "_1", "0", "1" means the left, current and right words' feature; "W" represent the string of word; "P" represent POS tag; "T" represent mention tag; "S" represent the binary-value single character feature.

Five best transformation rules are illustrated in Table 2. For example, MR3 means "if current word's POS tag is *nt*, then it is a *ORG*". Following example well describe the process of applying these rules.

俄罗斯/new/Russia 法院/nt/court

→俄罗斯/new/Russia $[$法院/nt/court$]_{ORG}$ **(MR3)**

→$[$ 俄 罗 斯 /new/Russia$]_{GPE}$ $[$ 法 院 /nt/court$]_{ORG}$ **(MR5)**

5 Entity Tracking

In our entity tracking/coreference model, the initial state is let each mention in a document form an entity, as shown in Figure 2 (a). And the objective function directs the learning process to insert or remove chains between mentions (Figure 2 b and c) to approach the goal state (Figure 2 f).

A list of rules is learned in greedy fashion, according to the objective function. When no rule that improves the current state of the training set beyond a pre-set threshold can be found, the training phrase ends. The objective function in our system is driven by the correctness of the binary classification for pair-wise mention pairs.

The TBL entity tracking model has more widely clustering/searching space as compare with previous strategies (Soon et al. 2001; Ng and Cardie, 2002; Luo et al., 2004). For example, the state shown in Figure 2 (d) is not reachable for them. Because they assume one mention should refer to its most confidential mentions or entities that before it, while A and B are obviously not in same entity, as we can see in Figure 2 (d). Thus C can refer to either A or B, but not both. While in TBL model, this state is allowed.

In order to keep our system robust, the transformation templates refer to only six types of simple features, as described below.

All these features do not need any high level tools (i.e. syntactic parser) and little external knowledge base. In fact, only a country name abbreviation list (171 entrances) and a Chinese

province alias list (34 entrances) are used to detect "alias" relation for String Match feature.

String Match feature (STRM): Its possible values are *exact*, *alias*, *abbr*, *left*, *right*, *other*. If two mentions are exact string matched, then return *exact*; else if one mention is an alias of the other, then return *alias*; else if one mention is the abbreviation of the other, then return *abbr*; else if one mention is the left substring of the other, then return *left*; else if one mention is the right substring of the other, then return *right*; else return *other*.

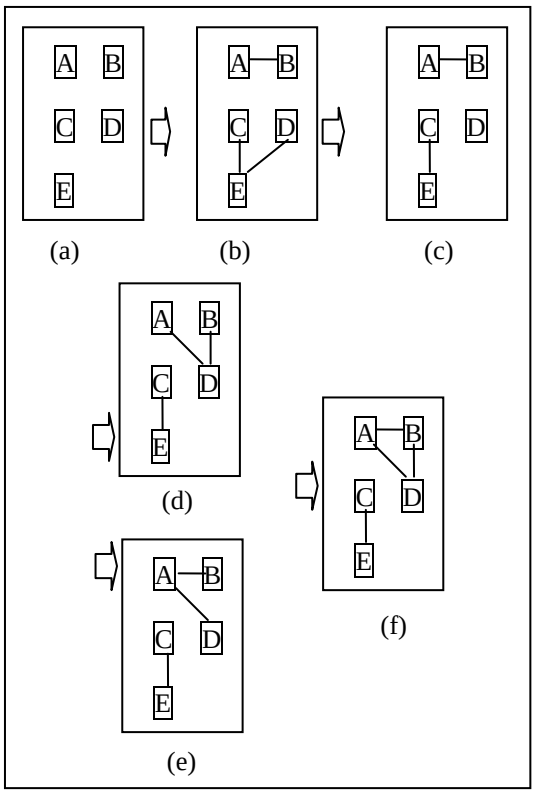

Figure 2. The procedure of TBL entity tracking/coreference model

Edit Distance feature I (ED1): Its possible values are *true* or *false*. If the edit distance of the two mentions are less than or equal to 1, then return *true*, else return *false*.

Token Distance feature I (TD1): Its possible values are *true* or *false*. If the edit distance of the two mentions are less than or equal to 1(i.e., there are not more than one token between the two mentions), then return *true*, else return *false*.

Mention Type (MT): Its possible values are *NAME*, *NOMINAL*, or *PRONOUN*..

Entity Type (ET): Its possible values are *PER*, *GPE*, *ORG*, *LOC*, or *FAC*.

Mention String (M): Its possible values are the actual mention string.

These six features can be divided into two categories: mention pair features (the first three) and single mention features (the other three). And the single mention features are suffixed with "L" or "R" to differentiate for left or right mentions (i.e. ETL represent the left mention's entity type).

Based on the six kinds of basic features, four simple transformation templates are used in our system, as listed in table 3.

Table 3. Templates for coreference model.

CT1: MTL,MTR,STRM
CT2: MTL,MTR,ETL,ETR,ED1
CT3: MTL,MTR,ETL,ETR,TD1
CT4: MTL,MTR,ML,MR

Table 4. Examples of transformation rules of coreference model.

CR1:CT1 NAME NAME EXACT LINK
CR2:CT2 NOMINAL NAME PER PER 1 LINK
CR3:CT1 NAME NAME ALIAS LINK
CR4:CT1 PRONOUN PRONOUN EXACT LINK

Though trained on different data set will learn different rules, the four rules listed in table 4 is the best rules that always been learned. For example, the first rule means that "**If** two *NAME* mentions are *exact* string matched, **then** insert a chain between them". The following example illustrates the process.

[美国/US]$_{GPE}$ 墩促[俄罗斯/Russia]$_{GPE}$ 基于人道原因释放 [美国/US]$_{GPE}$[商人/businessman]$_{NPER}$[波普/Bopu]$_{PER}$

→ [美国/US]$_{GPE-1}$ 墩促[俄罗斯/Russia]$_{GPE}$ 基于人道原因释放[美国/US]$_{GPE-1}$[商人/businessman]$_{NPER}$[波普/Bopu]$_{PER}$ **(CR1)**

→ [美国/US]$_{GPE-1}$ 墩促[俄罗斯/Russia]$_{GPE}$ 基于人道原因释放[美国/US]$_{GPE-1}$[商人/businessman]$_{NPER-2}$[波普/Bopu]$_{PER-2}$ **(CR2)**

6 Feedback

There are three reasons push us apply feedback technique in the EDT system. The first is to determine whether a signal character is an abbreviation is discourse depended. For example, Chinese character "中" can represents both a country name "China" and a common preposition "in". If it can links to "中国/China" by coreference model, it is likely to represent

"China". The second is the definition of mentions is hard to hold, especially the nominal mentions. An isolated mention is more likely not to be a mention. The third is to pick up lost mention according to its multi-appearance in the discourse. In fact, [Ji and Crishman, 2004] has used five hubristic rules based on coreference results to improve the name recognition result. While in this section we will present an automatic method.

The feedback technique is employed by using entity features in mention detection model. In our model, the transformation templates refer to the number of mentions in the entity, the single character feature, the entity type feature, the mention type feature and mention string, as listed follows.

SDD: Its possible values are the combination of the mention type and entity type of the mention string in discourse: *PER*, *GPE*, *ORG*, *LOC*, *FAC*, *NPER* (NOMINAL PER), *NGPE*, *NORG*, *NLOC*, *NFAC*, *PPER* (PROUNOUN PER), *PGPE*, *PORG*, *PLOC*, and *PFAC*.

SC2, SC3, SC4: Their possible values are *true* or *false*. If the word string appear not less than 2 (3, 4) times in the discourse then return *true*, else return *false*.

PDD: presents the combination of the mention type and entity type of the mention in discourse. Its possible values are same with "SDD".

PC2: Its possible values are *true* or *false*. If the mention belong to an entity has not less than 2 mentions then return *true*, else return *false*.

S0: Its possible values are *true* or *false*. If the mention is a single character word then return *true*, else return *false*.

W0: string of the mention.

Table 5. Templates for feedback.

FT1: SDD,SC2	**FT4:** PDD,PC2,S0
FT2: SDD,SC3	**FT5:** PDD,PC2,S0
FT3: SDD,SC4	**FT6:** PDD,PC2,W0

Table 6. Examples of transformation rules of feedback.

FR1: FT1 PER T PER	**FR4:** FT4 NORG F O
FR2: FT5 GPE F 1 O	**FR5:** FT3 PGPE F O
FR3: FT4 NFAC F O	

The first rule means that "if a word in the document appears as person name more than two times, then it is a person name". This rule can pick up lost person names. The second rule means that "if a GPE mention is isolated and it is a single character word, then it is not a mention". This rule can throw away isolated abbreviation of GPE, as illustrated in the following example.

…[波普/Bopu]$_{PER-3}$ 在星期三被[俄罗斯/Russia]$_{GPE-2}$[法院/court]$_{ORG-4}$[**以/by**]$_{GPE-6}$ 间谍罪判处 20 年徒刑 …

→…[波普/Bopu]$_{PER-3}$ 在星期三被[俄罗斯/Russia]$_{GPE-2}$[法院/court]$_{ORG-4}$ **以/by** 间谍罪判处 20 年徒刑 … **(FR2)**

7 Experiments

Our experiments are conducted on Chinese EDT corpus for ACE project from LDC. This corpus is the training data for ACE evaluation 2003. The corpus has two types, paper news (nwire) and broadcast news (bnews). the statistics of the corpus is shown in Table 7.

Table 7. Statistics of the ACE corpus.

	nwire	bnews
Document	99	122
Character	55,000	45,000
Entity	2517	2050
Mention	5423	4506

Because the test data for ACE evaluation is not public, we randomly and equally divide the corpus into 3 subsets: set0, set1, set2. Each consists of about 73 documents and 33K Chinese Characters[2]. Cross experiments are conducted on these data sets. ACE-value is used to evaluate the EDT system; and precision (P), recall (R) and F (F=2*P*R/(P+R)) to evaluate the mention detection result.

In the experiments, we first use one data set train the mention detection system; then use another set train the coreference model based on the output of the mention detection; finally use the other set test. In practice, we can retrain the mention detection model use the two train set to get higher performance.

Table 8. EDT and mention detection results.

Method	EDT	Mention Detection		
	ACE-value	R	P	F
Tag	55.7±1.6	62.3±1.0	85.0±1.4	71.9±0.6
SegTag	61.6±3.6	70.9±4.5	81.9±1.0	75.9±2.6
SegTag+F	63.3±2.0	68.0±4.8	83.8±1.2	75.0±3.1

[2] Two of the documents (CTS20001110.1300.0506, and XIN20001102.2000.0207) in the corpus are not use for serious annotation error.

In Table 8, "SegTag" represent the mention detection system integrated with segmentation adaptation, "Tag" represent the mention detection system without segmentation adaptation. "+F" means with feedback.

The ACE-value of our Chinese EDT system is better than 58.8% of Florian et al. (2004). In fact, the two systems are not comparable for not basing on the same training and test data. However both corpora are under the same standard from ACE project, and our training data (about 66K) is smaller than Florian et al. (2004) (about 80K). Therefore, it is an encouraging result.

Segmentation adapting and feedback can improve 7.5% of ACE-value for the whole system. As we can see from Table 8, using TBL method to adapt standard or correct errors can improve the mention detection performance especially recall, and word segmentation adapting is essential for mention detection. Feedback can improve the precision of mention detection with loss of recall. The two techniques can significantly improve the EDT performance, since the p-value of the T-test for the performance of "SegTag" to "Tag" is 96.7%, while for "SegTag+F" to "Tag" is 98.9%. The recall of mention detection is dropped after feedback because of the great effect of rule **FR**2, 3, 4 and 5 as illustrated in table 6.

8 Conclusion

In this paper, we integrate the mention detection model and entity tracking/coreference model into a unified TBL framework. Experimental results show segmentation adapting and feedback can significantly improve the performance of EDT system. And even with very limited knowledge and shallow NLP tools, our method can reach comparable performance with related work.

References

Eric Brill. 1995. Transformation-based error-driven learning and natural language processing: a case study in Part-of-Speech tagging. In: *Computational Lingusitics*, 21(4).

R Florian, H Hassan, A Ittycheriah, H Jing, N Kambhatla, X Luo, N Nicolov, and S Roukos. 2004. A statistical model for multilingual entity detection and tracking. In *Proc. of HLT/NAACL-04*, pages 1-8, Boston Massachusetts, USA.

Jianfeng Gao, Mu Li and Changning Huang. 2003. Improved souce-channel model for Chinese word segmentation. In *Proc. of ACL2003*.

Jianfeng Gao, Andi Wu, Mu Li, Changning Huang, Hongqiao Li, Xinsong and Xia, Haowei Qin. 2004. Adaptive Chinese word segmentation. In *Proc. of ACL2004*.

Niyu Ge, John Hale, and Eugene Charniak. 1998. A statistical approach to anaphora resolution. In *Proc. of the Sixth Workshop on Very Large Corpora*.

Sanda M. Harabagiu, Razvan C. Bunescu, and Steven J. Maiorano. 2001. Text and knowledge mining for coreference resolution. In *Proc. of NAACL*.

J. Hobbs. 1976. Pronoun resolution. Technical report, Dept. of Computer Science, CUNY, Technical Report TR76-1.

A Ittycheriah, L Lita, N Kambhatla, N Nicolov, S Roukos, and M Stys. 2003. Identifying and tracking entity mentions in maximum entropy framework. In *HLT-NAACL 2003*.

Heng Ji and Ralph Grishman. 2004. Applying Coreference to Improve Name Recognition. In *ACL04 Reference Resolution and its Application Workshop*.

Xiaoqiang Luo, A. Ittycheriah, H. Jing, N. Kambhatla, S. Roukos.2004. A Mention-Synchronous Coreference Resolution Aogorithm Based on the Bell Tree. In *Proc. of ACL2004*.

R. Mitkov. 1998. Robust pronoun resolution with limited knowledge. In *Proc. of the 17th International Conference on Computational Linguistics*, pages 869-875.

MUC. 1996. *Proceedings of the Sixth Message Understanding Conference* (*MUC*-6). Morgan Kaufmann, San Mateo, CA.

NIST. 2003. The ACE evaluation plan. www.nist.gov/speech/tests/ace/index.htm.

Wee Meng Soon, Hwee Tou Ng, and Chung Yong Lim. 2001. A machine learning approach to coreference resolution of noun phrases. *Computational Linguistics*, 27(4):521-544.

M. Vilain, J. Burger, J. Aberdeen, D. Connolly and L. Hirschman. 1995. A Model-Theoretic coreference scoring scheme. In *Proc. of MUC-6*, page45-52. Morgan Kaufmann.

Xiaofeng Yang, Guodong Zhou, Jian Su, and Chew Lim Tan. 2003. Coreference resolution using competition learning approach. In *Proc. of ACL2003*.