Lang Resources & Evaluation (2008) 42:197–217
DOI 10.1007/s10579-008-9066-8

Product named entity recognition in Chinese text

Jun Zhao · Feifan Liu

Published online: 17 April 2008

Abstract There are many expressive and structural differences between product names and general named entities such as person names, location names and organization names. To date, there has been little research on product named entity recognition (NER), which is crucial and valuable for information extraction in the field of market intelligence. This paper focuses on product NER (PRO NER) in Chinese text. First, we describe our efforts on data annotation, including well-defined specifications, data analysis and development of a corpus with annotated product named entities. Second, a hierarchical hidden Markov model-based approach to PRO NER is proposed and evaluated. Extensive experiments show that the proposed method outperforms the cascaded maximum entropy model and obtains promising results on the data sets of two different electronic product domains (digital and cell phone).

Keywords Information extraction · Product named entity recognition · Hierarchical hidden Markov model

This research was conducted under the framework of the Chinese Linguistic Data Consortium (ChineseLDC). In the first phase, ChineseLDC created a series of fundamental Chinese language resources, including Comprehensive Chinese Lexicon, Chinese Grammatical Knowledge Base (frequent words), Word-segmented and POS-tagged Chinese Corpus, Syntactic Treebank, Chinese–English Parallel Corpus, Chinese Semantic Lexicon, etc. Construction of the Product Named Entity Tagged Corpus and development of the Automatic Product Named Entity Recognition Tool are among the tasks of the second phase of ChineseLDC.

J. Zhao (✉) · F. Liu
National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100080, China
e-mail: jzhao@nlpr.ia.ac.cn

F. Liu
e-mail: ffliu@nlpr.ia.ac.cn

1 Introduction

Named entity recognition (NER) plays an important role in information extraction (IE) and many other applications. Previous research on NER falls mainly into two categories. One is general NER aiming to recognize person (PER), location (LOC), organization (ORG), time (TIM) and numeral (NUM) expressions mostly in the news domain, and the other is to identify some domain-specific proper names such as genes and proteins in biology. However, to our knowledge, there is little prior research on product NER (PRO NER), which is crucial and valuable for IE in the field of market intelligence. There are many expressive and structural differences between a product named entity (PRO NE) and general named entity (NE).[1] The paper focuses on PRO NER in Chinese text. The main contributions are as follows:

- *Establishment of PRO NE annotation specifications and construction of an annotated corpus*: Based on large-scale text from the Internet, we defined three types of PRO NEs and thoroughly analyzed their characteristics. Furthermore, several PRO NE annotation specifications were established and the first manually annotated corpus for PRO NER in Chinese text was constructed.
- *New research findings on methods of automatic PRO NER in Chinese text*: Because PRO NEs often have complex structures and flexible expressions, a hierarchical hidden Markov model (HHMM) (Fine et al. 1998) based approach for PRO NER is proposed. In this approach, two HHMMs are established using word form features and part of speech (POS) features, respectively. The two HHMMs are combined with knowledge base and heuristics to utilize diverse contextual features. The experiments show that the proposed method outperforms the cascaded maximum entropy (ME) model and obtains promising results in the electronic digital domain and cell phone domain.

2 Related work

Up to now, not much work has been done on PRO NER. Pierre (2002) developed an English NER system capable of identifying product names in product reviews. It employed a simple Boolean classifier for identifying product names, which is similar to token matching and is not applicable for PRO NER because of its more flexible and variant expressions. Bick (2004) recognized NEs including product names based on a constraint-grammar-based parser for Danish. This rule-based approach is highly dependent on the performance of the Danish parser. Niu et al. (2003) presented a bootstrapping approach for English NER using two successive learners [parsing-based decision list and hidden Markov model (HMM)], which produced promising experimental results (F-measure: 69.8%) on PRO NEs. The main advantage of this method is that manual annotation of a sizable training corpus can be avoided, but it

[1] For purposes of clarity and precision, the singular forms "product named entity" and "named entity" are abbreviated "PRO NE" and "NE", respectively, while the plural forms "product named entities" and "named entities" are abbreviated "PRO NEs" and "NEs," respectively.

suffers from two problems: it is difficult to find sufficient concept-based seeds for bootstrapping; and it is highly dependent on parser performance.

Research on PRO NER is still in an early stage, especially in Chinese free texts. There is neither a systematic specification for PRO NE tagging nor a manually tagged corpus for studying automatic PRO NER. However, a considerable amount of work has been done in the last decade on the general NER task and biological NER task. The typical machine learning based approaches for English NER include transform-based learning (Aberdeen et al. 1995), the HMM (Bikel et al. 1997; Collier et al. 2000), the ME model (Borthwick 1999), support vector machine learning (Yi et al. 2004), the decision tree model (Sekine et al. 1998), etc. For research on Chinese NER, the prevailing methods are also machine learning-based approaches, combined with knowledge bases or heuristic rules. In short, in the field of NER, researchers have tried to use hybrid statistical models that can combine different feature types at different levels and integrate some heuristics and external knowledge bases as well.

In this paper, we first propose a systematic specification for PRO NE tagging, by which a sizable corpus is built with manually annotated PRO NEs. Second, we present a hybrid approach based on the HHMM for Chinese PRO NER and conduct extensive experiments for evaluation.

3 The construction of a PRO NE tagged corpus

In this section, we describe our efforts on data annotation, including well-defined specifications, data analysis, and development of a corpus with annotated product named entities.

3.1 The definition

It is difficult to precisely define what kinds of expressions should be considered as PRO NEs. Generally, a PRO NE contains a twofold meaning. On one hand, it must be an expression referring to a determinate product category. On the other hand, it must indicate "named" information.

Based on the analysis of large-scale real text, we found that PRO NEs have the following characteristics. First, they are often composed of product brands and product types, which contain important and discriminative PRO NE information and distinguish this kind of NE from others. We call them *the basic elements of PRO NE*. Second, in some cases, there are embedded expressions describing the attributes and categories of products. We call them *the complementary elements of PRO NE*. For example, in the PRO NE "摩托罗拉 (Motorola) V8088 折叠 (clamshell) 手机 (cell phone)", "摩托罗拉 (Motorola)" is a product brand, "V8088" is a product type, "手机 (cell phone)" is the category word of a kind of products, and "折叠 (clamshell)" is a word describing an attribute of this kind of products. In real contexts, however, some of these elements can often be omitted when referring to a PRO NE.

From the above observation, we believe that a nominal expression must satisfy the following prerequisite in order to be considered as a PRO NE in text.

It contains either a brand name or a type name, or both of them.

For example, "爱国者 (AIGO) 闪存 (USB Flash Drive)" is a PRO NE, while "数码 (digital) 相机 (camera) 产品 (product)" is not since "digital camera product" has no named information; "EasyShare 系列 (series) 数码 (digital) 相机 (camera)" is a PRO NE, while "智能型 (intelligent) 手机 (cell phone)" is not since "EasyShare" is a specific series of Kodak brand, while "intelligent" is a common attribute of many brands of cell phones.

3.2 The tagging set

The tagging set of PRO NE includes three tags, namely Brand Name, Product Type and Product Name, which are defined as follows:

Brand Name (*BRA*) refers to the proper name of a product trademark, such as "明基 (BenQ)" in Example 1.
Product Type (*TYP*) indicates the version or series information of a product, which can consist of numbers, Latin letters, or other symbols such as "+" and "−". In Example 2, "*Pro90IS*" is a TYP.
Product Name (PRO) indicates a PRO NE in text, which can be composed of the Brand Name, the Product Type, the category word of a PRO NE, and expressions describing the attributes of a product. Not all of them are absolutely necessary. In Example 2, "*Canon 334万 (3.34 million) 像素 (pixels) 数码 (digital) 相机 (camera) Pro90IS*" is a PRO.

Among them, BRA and TYP are often nested inside PRO, such as in Example 2.

Example 1 明基 (BenQ)/BRA 的 (of) 市场占有率 (market share) 稳步 (steadily) 上升 (rise) 。 (.)
(BenQ's market share is rising steadily.)

Example 2 公司 (The company) 即将 (will soon) 推出 (release) [Canon/BRA 334 万 (3.34 million) 像素 (pixels) 数码 (digital) 相机 (camera) Pro90IS/TYP]/PRO 。 (.)
(The company will soon release the Canon Pro90IS 3.34-megapixel digital camera.)

3.3 The specification of tagging PRO NEs

Based on the definition of three types of PRO NEs and their characteristics, we established several PRO NE annotation specifications.

3.3.1 Main principles

Some principles should be followed in the process of manual tagging of PRO NEs.

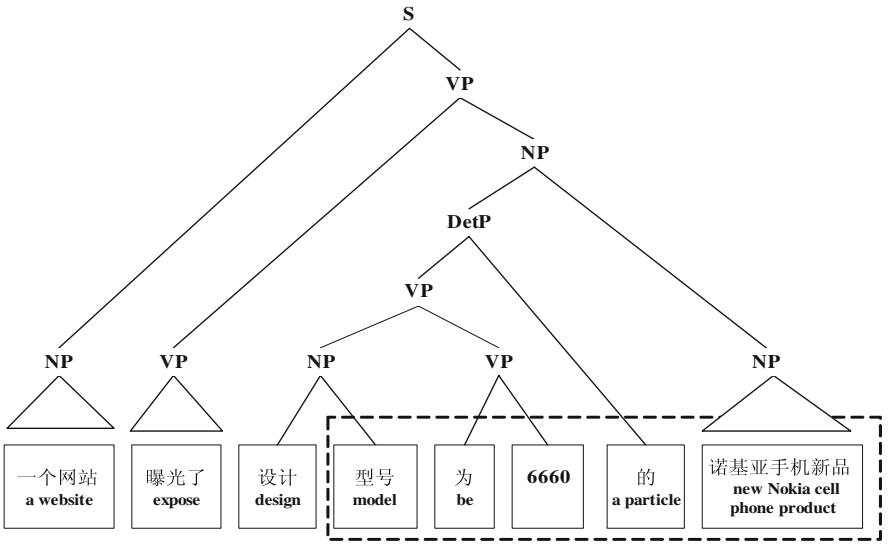

Fig. 1 An example to illustrate the annotation of PRO NEs

First, the tagged PRO NEs should have determinate and relatively self-contained meaning. In other words, the tagged unit should refer to a determinate PRO NE. As in one of the aforementioned examples, we should tag "EasyShare 系列 (series) 数码 (digital) 相机 (camera)" as a PRO NE even though it has no brand elements, because "EasyShare" carries distinguishable information of the Kodak brand.

Second, the annotation of a PRO NE cannot destroy the sound syntactic structure of the sentence. For example, the syntactic tree of Example 3 (which is created based on Chinese grammar) is shown in Fig. 1, where "设计 (design) 型号 (model)" and "为 (be) 6660/TYP" create a VP, which is combined with "的 (a particle)" to create a DetP, which is further combined with "诺基亚 (Nokia)/BRA 手机 (cell phone) 新品 (new product)" to create an NP. If [型号 (model) 为 (be) 6660/TYP 的 (a particle) 诺基亚 (Nokia)/BRA 手机 (cell phone) 新品 (new product)] is tagged as a PRO NE only based on the fact that it has already covered the basic elements discussed above, then the syntactic structure of the sentence will be destroyed. Therefore, in this case, [设计 (design) 型号 (model) 为 (be) 6660/TYP 的 (a particle) 诺基亚 (Nokia)/BRA 手机 (cell phone) 新品 (new product)] should be tagged as a PRO NE entirely. We have no parsing reference for each sentence, but with this principle the native annotators still can effectively avoid reducing the readability of the sentence structure, which also leads to more consistent annotation results.

Example 3 昨天 (Yesterday), (,) 某网站 (a website) 曝光了 (expose) [设计 (design) 型号 (model) 为 (be) 6660/TYP 的 (a particle) 诺基亚 (Nokia)/BRA 手机 (cell phone) 新品 (new product)]/PRO 。(.)
(A new Nokia cell phone, design model 6660, was exposed on a website yesterday.)

3.3.2 The specification

Based on the above principles, in order to improve the consistency of the manual annotation process, we constitute more detailed specifications. Note that in example sentences hereafter, the tag just after the slash is either the POS tag or the PRO NE tag for the word (or punctuation) just before the slash. The POS tag set is listed in the Appendix and the PRO NE tag consists of BRA, TYP and PRO as defined in Sect. 3.2.

3.3.2.1 The tagging of quotation marks attached to a PRO NE In a PRO NE, quotation marks are used to set off alias names, series names or brand names. They can be tagged as follows:

- Quotation marks which are used to set off alias names or series names are tagged inside the PRO NE.

Example 4 这 (This)/r 就是 (is)/v ["/w 蓝 (blue)/a 精灵 (eidolon)/n "/w]/PRO 的 (of)/u 主体 (main part)/n 。 (.)/w
(This is the main part of the "Blue Eidolon".)

- Quotation marks which are used to set off brand names are tagged inside the PRO NE.

Example 5 配备 (equipped with)/v 了 (an auxiliary word)/u ["/w 森海塞尔 (Sennheiser)/BRA "/w 耳机 (headphones)/n]/PRO 的 (of)/u 产品 (product)/n 。 (.)/w
(a product equipped with "Sennheiser" headphones)

3.3.2.2 The tagging of a Chinese brand name and its English translation equivalent Sometimes, a PRO NE contains both a Chinese brand name and its English translation equivalent. In such cases, they can be tagged as follows:

- When both a Chinese brand name and its English translation equivalent appear inside a PRO NE, if there are no other words, characters or symbols between them, they are tagged as a BRA.

Example 6 [明基 BenQ/BRA M770GT/TYP 手机 (cell phone)/n]/PRO
(明基 BenQ M770GT cell phone)

- When both a Chinese brand name and its English translation equivalent occur inside a PRO NE, if there is only a simple conjunctive symbol between them, they are tagged as a BRA.

Example 7 [DOGGY-刀客/BRA MP3/nx 随声听 (personal stereo)/n]/PRO
(DOGGY-刀客 MP3 personal stereo)

3.3.2.3 The tagging of expressions like "…系列 (series)" and "…型 (type)" When expressions like "…系列 (series)" and "…型 (type)" are contained in a PRO NE, they can be viewed as the extension of TYP and are included inside PRO.

Example 8 [摩托罗拉 (Motorola)/BRA A系列 (Series)]/PRO
(Motorola A series)

3.3.2.4 The tagging of coordinate structures and elliptical structures Sometimes, coordinate structures and elliptical structures appear inside a PRO NE. In such cases, they can be tagged as follows:

- If two expressions each describing a PRO NE are connected by a conjunction, they are tagged separately as two PRO NEs.

Example 9 三星 (Samsung)/BRA 的(of)/u [X100/TYP]/PRO 和 (and)/c [X600/TYP]/PRO 以及 (and)/c 西门子 (Siemens)/BRA 的 (of)/u [C60/TYP]/PRO 和 (and)/c [MC60/TYP]/PRO
(Samsung X100 and X600 and Siemens C60 and MC60)

In the above example, X100 and X600 are two kinds of Samsung products, and they are connected by the conjunction "and". We tag them separately as two PRO NEs. Likewise, C60 and MC60 are tagged separately as well.

- In some cases, some common components (usually the basic elements) of the coordinate structures of PRO NEs are omitted. The following rules are followed in such cases.
  - Conjunctions, "、" (a Chinese punctuation mark used to separate items in a list) and commas are not tagged inside PRO.
  - In order to retain the pragmatic function of a PRO NE, this type of expression is tagged as two separate PRO.
  - The tagging processes cannot invalidate the syntactic structure of the sentence, except the coordinate structure.

Example 10 [EOSDCS3/TYP 型 (type)/k]/PRO 、 /w [EOS–IN/TYP 型 (type)/k 相机 (camera)/n]/PRO 的 (of)/u 外观 (exterior)/n 设计 (design)/n
(the exterior design of the EOSDCS3 and EOS-IN cameras)

In Example 10, "EOSDCS3 型 (type)、EOS–lN 型 (type) 相机 (camera)" is a coordinate structure, where "EOSDCS3 型 (type)" is an elliptical structure. We tag two separate PRO as "[EOSDCS3/TYP 型 (type)/k]/PRO" and "[EOS–lN/TYP 型 (type)/k 相机 (camera)/n]/PRO" and exclude the "、" in the PRO NE annotation.

3.3.2.5 Annotating to the maximum length of possible extension When tagging PRO NEs, we follow the rule of annotating to the maximum length of possible extension.

- The basic elements of a PRO NE, the descriptive modifiers embedded inside the basic elements (like "胶卷 (film)" in Example 11), and the circumjacent modifiers which describe the inherent attributes of a PRO NE (like "超薄 (super-thin)

钛金属 (titanium)" in Example 12), especially the modifiers containing special information about style, design and pattern, can all be included in a PRO if they form an agglutinate structure. Quotation marks and brackets are also allowed inside a PRO.

Example 11 [柯达 (Kodak)/BRA DCS520/TYP 数码 (Digital)/n 单反 (SLR)/b]/PRO, /w 采用 (adopt)/v 当前 (current) [佳能 (Canon)/BRA EOS/nx 系列 (series)/q 胶卷 (film)/n 相机 (camera)/n]/PRO 中 (in)/j [顶级 (most superior)/b 专业 (professional)/n 机型 (model)/n EOS–lN/TYP]/PRO 。 (.)/w
(The Kodak DCS520 Digital SLR (single-lens reflex) camera adopts the most superior professional model, EOS-lN, of Canon's current EOS series of film cameras.)

Example 12 [超薄 (super-thin)/b 钛金属 (titanium)/n 手机 (cell phone)/n BenQ/BRA M770GT/TYP]/PRO
(super-thin titanium cell phone BenQ M770GT)

- For two appositives, if there is no conjunctive symbol between them, then they are tagged as a single PRO; otherwise, they are separately tagged. In Example 13, there is a conjunctive symbol "和 (and)" between two appositives, ""/w 哈Q族 (HaQZu)/Ng "/w Q268/TYP" and ""/w 幻影 (apparition)/n "/w Q800/TYP", so they are tagged as two PRO.

Example 13 波导 (Bird)/BRA 的 (of)/u 两 (two)/NUM 款 (a quantifier)/q 新 (new) 机 (types)/n ["/w 哈Q族 (HaQZu)/Ng" /w Q268/TYP]/PRO 和 (and)/c ["/w 幻影 (Apparition)/n" /w Q800/TYP]/PRO
(two new types of Bird cellphones "HaQZu" Q268 and "Apparition" Q800)

- The "maximum length" rule must comply with the main principle of maintaining the validity of the sentence's syntactic structure. In Example 14, if we tag "Coolpix 4200 和 (and) Coolpix 5200" as a single PRO, the syntactic structure of the sentence would be destroyed, so we tag them seperately.

Example 14 Nikon/ORG 发布 (announce)/v 400万 (SD400)/NUM 以及 (and)/c 500万 (SD500)/NUM 数码 (digital)/n 相机 (camera)/n [Coolpix 4200/TYP]/PRO 和 (and)/c [Coolpix 5200/TYP]/PRO
(Nikon announces its SD400 and SD500 digital cameras, Coolpix 4200 and Coolpix 5200)

3.3.2.6 TYP Annotation *Product Type (TYP)* is usually composed of numbers, letters, and other symbols. If they are expressions (usually in English) about the version or series information of a product, then they can be combined and tagged as a single TYP. However, Chinese characters are not considered to be a TYP, nor subpart of TYP, although some of them do contain version or series information. For instance, in "2005 新年贺岁 (Happy New Year) 版 (version) 手机 (cell phone)", "新年贺岁 (Happy New Year) 版 (version)" is not considered to be a TYP.

3.3.2.7 BRA Annotation In some cases, it is very difficult to determine whether an expression is an organization name or a brand name. In such cases, we tag the expression as an organization.

Example 15 这 (This)/r 是 (is)/v 因为 (because)/p 三星 (Samsung)/ORG 一贯以来 (consistently)/d 精细的 (fine)/a 做工 (workmanship)/n 和 (and)/c 时尚的 (fashionable)/n 设计 (design)/n 。 (.)/w
(This is because of Samsung's consistently fine workmanship and fashionable designs.)

In Example 15, "三星 (Samsung)" may refer to Samsung Corporation or the Samsung brand, and thus we tag it as an organization. In comparison, in Example 11, "柯达 (Kodak)" refers to the Kodak brand because it co-occurs with "DCS520 数码 (Digital) 单反 (SLR)", so we tag it as BRA.

3.4 Construction of the CASIA_PRO corpus

We collected web pages related to product information, such as product releases, market trends, and product evaluations. These web pages were converted into plain text formats and all of them are non-structured free texts. These text files constitute the CASIA_PRO corpus. Currently, the size of CASIA_PRO1.2 is about 1,000,000 Chinese characters, including more than 1,500 web page texts in the fields of telecommunications and electronic digital equipment.

The corpus was processed pipeline through word segmentation, POS tagging, and general NER tagging (Wu et al. 2003). Then, the NE tags were proofread manually.

3.4.1 The manual annotation process of PRO NEs

Three students majoring in linguistics manually annotated PRO NEs in the corpus. The annotation process consists of three phases: pre-annotation, consistency testing and large-scale annotation.

- Pre-annotation: first of all, the three annotators studied the draft of the specification for PRO NE annotation. After that, we selected a portion of the CASIA_PRO corpus. The three annotators individually annotated the small corpus, found unreasonable items in the specification, and modified the specification draft accordingly. This process was repeated several times.
- Consistency testing: after the pre-annotation, we conducted consistency testing on the pre-annotated corpus. If the consistency reached a certain threshold, it meant that the specification met the needs of high-quality corpus annotation. Otherwise, the pre-annotation phase was resumed. The consistency testing method is discussed in detail in Sect. 3.4.2.
- Large-scale annotation: after we finalized the specification, the three annotators manually annotated the CASIA_PRO corpus.

3.4.2 Consistency testing: Kappa coefficient

To measure the annotation consistency, we conducted consistency testing experiments on a sample of the corpus (5% of the CASIA_PRO corpus which was annotated by the three annotators individually) and modified the annotation specification based on the testing results. We used a Kappa coefficient (Carletta 1996; Sigel et al. 1988) to measure the annotation consistency, which was computed as:

$$k = \frac{P(A) - P(E)}{1 - P(E)} \tag{1}$$

$$P(A) = \frac{\sum_i^n \delta(l_{i1}, l_{i2}, l_{i3})}{n} \tag{2}$$

$$P(E) = \sum_i^m \left(\frac{n_{i1}}{n} \cdot \frac{n_{i2}}{n} \cdot \frac{n_{i3}}{n}\right) \tag{3}$$

where $P(A)$ is the proportion of times that the annotators agree and $P(E)$ is the proportion of times that we would expect them to agree by chance; n is the total number of samples; n_{i1}, n_{i2}, n_{i3} are the total numbers of the samples that the first, second, and third annotator put into the ith category, respectively; m is the number of categories; and l_{i1}, l_{i2}, l_{i3} are the labels that the first, second, and third annotator assign to the ith sample, respectively. $\delta(l_{i1,} l_{i2,} l_{i3}) = 1$ only when $l_{i1} = l_{i2} = l_{i3}$; otherwise $\delta(l_{i1,} l_{i2,} l_{i3}) = 0$.

Our consistency testing experiments on the sample corpus achieved a Kappa coefficient of 0.81, which demonstrates that the consistency of PRO NE annotation is relatively satisfactory.

3.4.3 Some statistics of the PRO NE annotated corpus

The NE statistics of the CASIA_PRO1.2 corpus are shown in Table 1. In total, there are 12,432 PRO, 5,047 BRA, 10,606 TYP, 424 PER, 1,733 LOC, and 4,798 ORG in the corpus.

4 Automatic PRO NER

PRO NER involves the identification of product proper names in unconstrained text and their classification into different kinds of PRO NEs, namely PRO, TYP, and BRA in this paper.

Table 1 Statistics of CASIA_PRO1.2 corpus

PRO	BRA	TYP	PER	LOC	ORG
12,432	5,047	10,606	424	1,733	4,798

4.1 The difficulties of automatic PRO NER

In this section, we explore some particular characteristics of various PRO NEs to attain a clear understanding of the challenges in recognizing them. In comparison with general NEs, PRO NEs have their own special characteristics.

- For general NEs, there are some cues indicating their occurrence in the text. For example, "公司 (company)" is a cue for an organization name and often acts as the ending word of the name. These cues are very useful for general NER. In contrast, PRO NEs have few such cues around them, which makes it more difficult to trigger the PRO NER process and leads to more boundary ambiguities.
- There are many category ambiguities in PRO NER.
  - An expression can be a general NE, PRO NE, or just a common word, according to its context. For instance, "苹果 (apple)" can refer to a BRA, ORG, or just a kind of fruit; an English word such as "professional" can be a common word or a component of TYP; a digit string may be a TYP, a NUM, or a TIM.
  - Some category ambiguities related to PRO NEs are very difficult to distinguish, especially between BRA and ORG. For example, "这款手机采用了三星风格的设计 (this type of cell phone uses a Samsung-style design)." In the example, it is difficult to classify the highlighted part as a BRA or ORG.
- PRO NEs have more flexible forms. The same entity can be expressed in several different forms due to spelling variations, word permutations, etc. For example, "柯达DX7630数码相机 (Kodak DX7630 digital camera)" versus "柯达数码相机DX7630 (Kodak digital camera DX7630)."
- PRO NEs frequently have nested structures. More efforts must be made to identify such PRO NEs.

4.2 Hybrid approach for PRO NE recognition

Based on observations from the real data, there are three features we can use for this task. First, the components inside PRO NEs have certain characteristics. For example, many PRO NEs have an alphanumeric string inside, which denotes type or series information. Second, various contextual information can be very helpful in boundary detection and the classification process. Third, related knowledge bases, such as brand lists, can also provide helpful information for PRO NER. We try to take full advantage of these features in our strategy, which can be performed in two steps.

- First, the triggering step for detecting potential PRO NE candidates is very critical. If triggering conditions are loosely set, a lot of noise will be introduced. In contrast, if triggering conditions are set rigorously, the recall of PRO NER will be seriously reduced. We use both a knowledge base and some heuristics to trigger PRO NE candidates. The knowledge base we used includes a Chinese brand word list and an English brand word list. All heuristics are either domain-independent or can be easily acquired without much time-consuming manual editing.
- After triggering candidates, we try to make use of contextual information to determine whether a candidate should be tagged as a PRO, BRA, TYP, or other

name. For this step, we use a hybrid statistical model which can utilize the various features inside or outside a candidate. Furthermore, since nested structures are highly frequent and their lengths are quite variable in PRO NEs, we use the HHMM (Fine et al 1998) as the statistical model considering that it is more powerful to model the multiplicity of length scales and the recursive nature of sequences at different stochastic levels than to use other flat models such as the HMM and ME Markov Model (MEMM) (McCallum et al. 2000).

4.2.1 Overall workflow of PRO NER

The workflow of PRO NER is illustrated in Fig. 2, which includes the following three steps:

- *Preprocessing*: Word segmentation, POS tagging, and general NER are primarily conducted using our off-the-shelf SegNer2.0 toolkit (Wu et al. 2003) on input text.
- *Generating PRO NE candidates*: PRO NE as well as ORG (single-brand-word name) candidates are triggered and generated via a knowledge base (a list of Chinese and English brand words automatically obtained from the Internet or a training set). BRA (or ORG) and TYP are triggered by a brand word list and the Type Characteristic Class (TCC) shown in Table 2, respectively (Table 2 gives the TCCs, their tags and some examples, which may trigger TYP candidates). Then PRO is triggered by BRA and TYP candidates as well as some cue words indicating type information, such as "版" (version), "系列" (series), and "型" (type). Once triggered, the corresponding NE candidates are generated by binding the trigger word with its contexts. In this step, the model structure (topology) of the HHMM is dynamically constructed.
- *Disambiguating candidates*: The boundary ambiguity and classification ambiguity of the candidates are resolved simultaneously. The Viterbi algorithm is applied for finding the most-likely state sequences based on the HHMM topology. The HHMM for PRO NER is described at length in Sect. 4.2.2.

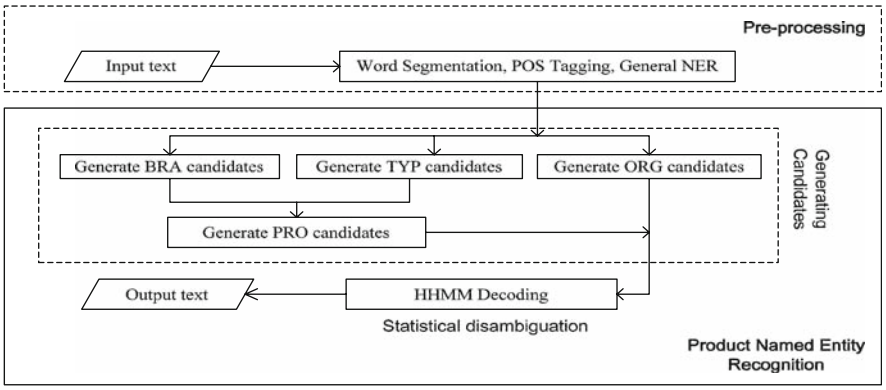

Fig. 2 Workflow for PRO NER

Table 2 TCC for TYP identification

TCC	Tags for TCC	Examples
Sequence of English letters	YZ	Powershot, Pro, i
Sequence of English letters and digits	ZS	T18、S100
Sequence of digits	SH	2100, 8088
Sequence of digits in single-byte character (SBC) case	QS	600
Sequence of English letters in SBC case	QZ	Pro
Other non-Chinese symbols	The original forms	@, -

4.2.2 HHMM for PRO NER application

The HHMM, a recursive hierarchical generalization of the HMM, is applied to address the PRO NER problem due to its ability of modeling the multiplicity of length scales and recursive nature of the sequences (Fine et al. 1998). By HHMM, PRO NER can be formulated as a tagging problem using Viterbi algorithm. Unlike the traditional HMM in POS tagging, the topology of the HHMM is not fixed and some states can be a similar stochastic model on themselves, which are called internal states, in contrast to production states which emit only observations.

For the HHMM-based PRO NER, the input sequence is a Chinese sentence which has been word-segmented, POS tagged, and general NE tagged. The sentence can be formalized as $w_1/t_1\ w_2/t_2 \ldots w_i/t_i \ldots w_n/t_n$, where w_i and t_i are the ith word and its part-of-speech (or general NE tag), respectively, and n is the number of words in the sentence. The POS tag set is the combination of the POS tag set from Peking University (PKU-POS shown in the Appendix) (Yu et al. 2003) and general NE (GNE) categories including PER, LOC, ORG, TIM, and NUM. We construct the HHMM for PRO NER as follows:

- the state set $\{S\}$, which consists of {GNE}, {BRA, PRO, TYP}, and {V}, where V is the vocabulary of Chinese words;
- the observation set $\{O\}$, which is equal to $\{V\}$.

In the above model, only PROs are internal states which may activate other production states such as BRA and TYP resulting in recursive HMM. Consistent with S. Fine's work, q_i^d $(1 \le \mathrm{d} \le \mathrm{D})$ is used to indicate the ith state in the dth level of the hierarchy. So, the PRO NER problem is to find the most-likely state activation sequence Q^*, a multi-scale list of states, based on the dynamic topology of the HMM given an observation sequence $W = w_1\, w_2 \ldots w_i \ldots w_n$, formulated as follows based on Bayes' rule (P(W)=1).

$$Q^* = \arg\max_Q P(Q|W) = \arg\max_Q P(Q)P(W|Q) \tag{4}$$

From the root node of the HHMM, the activation flows to all other nodes at different levels according to their transition probability. For description convenience, we take the kth level as an example (activated by the mth state at the $k-1$th level).

$$P(Q) \cong \underbrace{p(q_1^k|q_m^{k-1})}_{vertical\ \ transition} \overbrace{p(q_2^k|q_1^k)\prod_{j=3}^{|q^k|} p(q_j^k|q_{j-1}^k, q_{j-2}^k)}^{horizontal\ \ transition} \tag{5}$$

$$P(W|Q) = \begin{cases} \cong \prod_{j=1}^{|q_{PS}^k|} p([w_{q_j^k-begin} \cdots w_{q_j^k-end}]|q_j^k), \ if\ q_j^k \notin \{IS\} \\ activate\ other\ states\ recursively, \ if\ q_j^k \in \{IS\} \end{cases} \tag{6}$$

where $|q^k|$ is the number of all the states in the kth level; $|q_{PS}^k|$ is the number of the production states in the kth level; $w_{q_j^k-begin} \ldots w_{q_j^k-end}$ indicates the word sequence corresponding to the state q_j^k. For $p([w_{q_j^k-begin} \ldots w_{q_j^k-end}]|q_j^k)$, we have the following estimations:

- If $q_j^k \in \{\{GEN\}, \{V\}\}$, then we assume that the results of pre-processing are correct, that is

$$p([w_{q_j^k-begin} \cdots w_{q_j^k-end}]|q_j^k) = 1 \tag{7}$$

- If q_j^k = *BRA*, for simplification, we assign $p([w_{q_j^k-begin} \ldots w_{q_j^k-end}]|q_j^k)$ a constant value as in Eq. 8, because a brand word may generate not only a BRA candidate but also an ORG candidate.

$$p([w_{q_j^k-begin} \ldots w_{q_j^k-end}]|q_j^k = BRA) = p(q_1^{k+1}|q_j^k) = 0.5 \tag{8}$$

- If q_j^k = *TYP*, TCC defined in Table 2 are applied, i.e., the words associated with the current state are replaced with their TCC tags. Then we can compute the emission probability of this TYP production state as the following equation, within which $|q_j^k|$ is the length of observation sequence associated with the current state.

$$p([w_{q_j^k-begin} \ldots w_{q_j^k-end}]|q_j^k = TYP) \cong p(tc_1|begin)p(end|tc_{|q_j^k|}) \prod_{m=2}^{|q_j^k|} p(tc_m|tc_{m-1}) \tag{9}$$

- If q_j^k = *PRO*, because PRO is an internal state, production states in the $(k + 1)$th level will be activated by this internal state through Eq. 6, and the activation process will revert when arriving at an end state. Thus hierarchical computation is implemented.

Figure 3 uses a simple example to illustrate the process of state transition of the HHMM in the application in PRO NER. In the example, a tree-layer HHMM is used. The parameters in the above equations are estimated from the training set based on maximum likelihood estimation, where the parameters are smoothed using the deletion interpolation approach proposed by Jelinek and Mercer (1980).

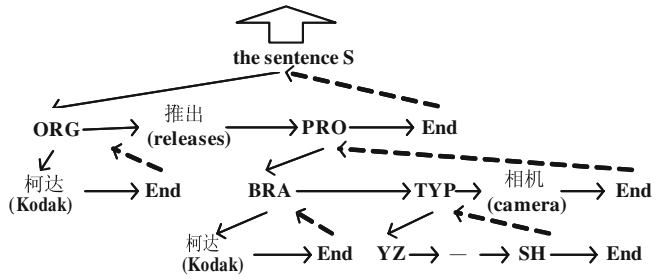

Fig. 3 A simple example to illustrate the process of state transition of the HHMM in the application in PRO NER

4.2.3 Integration of two HHMM instances

In Sect. 4.2.2, we have implemented a simple HHMM for PRO NER, which we call HHMM-1. Note that in HHMM-1, we exploit the contextual features only at levels of word forms and semantic categories (i.e., general NE types). In order to investigate the effect of POS information for PRO NER, we construct another HHMM (HHMM-2) based on POS tag information. After that, we hope to exploit multi-levels of contextual features for PRO NER by integrating HHMM-1 and HHMM-2.

The difference between HHMM-2 and HHMM-1 is the state set S_{II} and observation set O_{II}. HHMM-2 uses $T= t_1\ t_2 \ldots t_i \ldots t_n$ as the observation sequence, i.e., $O_{\mathrm{II}} = \{\mathrm{POS}\}$. Accordingly, $S_{\mathrm{II}} = \{\{POS\}, \{GNE\}, BRA, TYP, PRO\}$, among which PRO is an internal state. In HHMM-2, PRO NER is formulated as follows.

$$Q^* = \arg\max_{Q} P(Q|T) = \arg\max_{Q} P(Q)P(T|Q) \tag{10}$$

The description and computation of HHMM-2 is similar to HHMM-1. Both models make use of the semantic classification information from general NE tags, and word form features make HHMM-1 more discriminative, while POS features lead to the robustness of HHMM-2. Intuitively, the integration of these two models may be helpful for improving the performance of PRO NER by balancing robustness and discrimination, as Eq. 11 indicates:

$$(Q^*) = \arg\max_{Q} P(Q|W,\mathrm{T}) = \arg\max_{Q} P(Q)P(W|Q) \ \times \ [P(Q)P(T|Q)]^{\beta} \tag{11}$$

where β is a tuning parameter for adjusting the weight of two models. Instead of its proper form, we often use logarithmic form for convenience.

$$(Q^*) = \arg\max_{Q} \{\log(P(Q)) + \log(P(W|Q)) + \beta \times [\log(P(Q)) + \log(P(T|Q))]\} \tag{12}$$

Note that instead of trying to combining the word forms and POS features into one sophisticated HHMM, we just integrate the two models in a very simple and

effective way, which enables us to investigate the integrating influence on performance and explore different roles of those two feature types separately as well. In addition, the HHMM is a generative model and it is also not straightforward to create one HHMM which can combine different types of observation features as the discriminative model does. One possible way to do this is to combine the two types of features into one compound feature such as "word + POS" for each word, whereby it is not feasible to evaluate the two feature types separately. Another problem with that is more serious data sparseness resulting from features with more refined granularity.

5 Experiments and analysis

We conducted extensive experiments to see whether the proposed method was suitable for Chinese PRO NER. The data set, evaluation metric, experimental results and analysis are presented in this section.

5.1 Data set preparation

The training data and testing data are selected from CASIA-PRO1.2, which was introduced in Sect. 3.4. We randomly select 140 texts (digital 70, cell phone 70) as an open test set (OpenTestSet), the rest as a training set (TrainingSet), from which 160 texts are extracted as a closed test set (ClosedTestSet). The NE statistics in two test sets are presented in Table 3. We can see that, there are 1800 PRO, 803 BRA, 1364 TYP, 39 PER, 207 LOC, and 614 ORG in the open test set, and there are 1,553 PRO, 513 BRA, 1,296 TYP, 55 PER, 248 LOC, and 619 ORG in the closed test set.

5.2 Evaluation metric

Due to the characteristics of variant forms in PRO NEs, a soft evaluation method is applied in our experiments to make the evaluation more reasonable. The main idea is that we score recognized NEs from three aspects: detection, classification, and boundary.

- NEs which are detected, classified correctly, and bounded correctly are scored 1.0.
- NEs which are detected and classified correctly, but have boundary errors, should be given a discounted score, such as 0.8, 0.6, or 0.4, according to the

Table 3 Numbers of NE instances in test set

Test set	PRO	BRA	TYP	PER	LOC	ORG
OpenTestSet	1,800	803	1,364	39	207	614
ClosedTestSet	1,553	513	1,296	55	248	619

number of errors and the position (beginning or end boundary error). Discounted scores are determined empirically according to error analysis: 0.4 indicates two boundary errors, 0.8 means only one error for the beginning boundary, and 0.6 denotes only one error for the end boundary. For example, the recognition result in "一(a)款(style)和弦 (chord)彩屏 (color-screen)[手机 (cell phone)三星 (Samsung) S508]/PRO近期 (recently)上市 (come on the market)" should be given more credit than in "一(a)款(style)和弦 (chord)彩屏 (color-screen)手机 (cell phone)[三星 (Samsung) S508 近期 (recently)]/PRO上市 (come on the market)."

Traditional metrics of precision, recall and F-measure are employed in our evaluations using the following formulae:

$$precision = \frac{soft\ scores\ of\ correctly\ recognized\ NEs}{number\ of\ recognized\ NEs} \tag{13}$$

$$recall = \frac{soft\ scores\ of\ correctly\ recognized\ NEs}{number\ of\ NEs\ in\ answers} \tag{14}$$

$$F - measure = \frac{2 * recall * precision}{recall + precision} \tag{15}$$

5.3 Experimental results and analysis

5.3.1 Evaluation on the influence of β in the integrated model

In the integrated model denoted as Eq. 12, the β value reflects the different contribution of two individual models to the overall system performance. The larger the β value, the greater the contribution made by HHMM-2. Figure 4-6 illustrate the

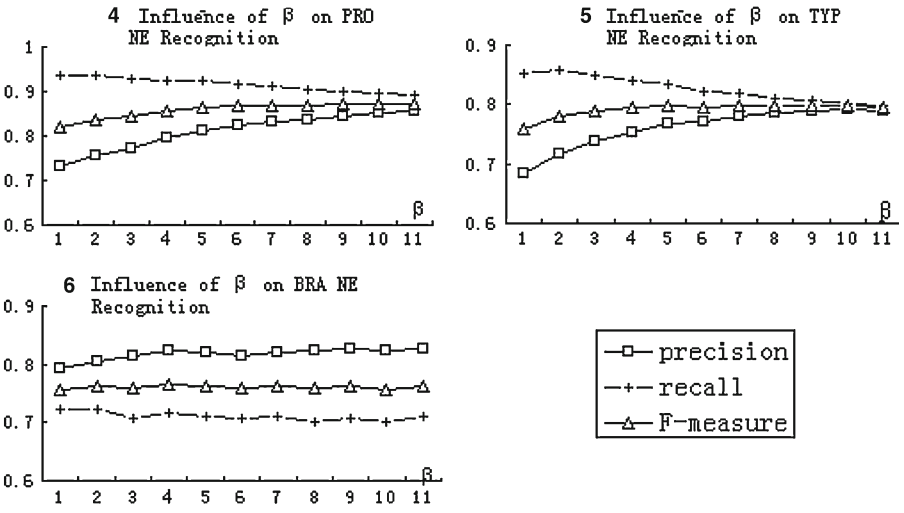

Fig. 4-6 Influence of β on PRO, TYP, and BRA NE recognition

varying curves of recognition performance with the β value on PRO, TYP, BRA, respectively.

Note that if β equals 1, then two models are integrated with equivalent weight. We can see that as β goes up, the F-measures of PRO and TYP first increase and then begin to decrease slightly after a period of flat growth. It can be explained that HHMM-2 mainly exploits POS and general NE features which can relieve the sparseness problem to some extent, which is more serious in HHMM-1 due to the lower level of contextual information such as word form. However, as β becomes larger, the problem of imprecise modeling in HHMM-2 becomes more salient and begins to show a side effect in the integrated model. Thus the performance can be improved at the early stage of β and finally declines. In contrast, the influence of β on BRA is negligible because its candidates are triggered by a relatively reliable knowledge base and its sub-model in the HHMM is assigned a constant as shown in Eq. 8.

From the performance curves, we can see that the integrated model can make up for the weakness of HHMM-1 (when $\beta = 0$) and achieve better performance on the whole. In addition, the performance improves as the β value increases, which indicates that HHMM-2 can make more contributions in the integrated model. This is due to the fact that high-level features are more robust since annotated data available is still limited at present. In our system, β is assigned a value of 8 based on the above experimental results.

5.3.2 Evaluation on the portability of PRO NER in two domains

We evaluate the performance of PRO NER in the corpus of the digital domain and cell phone domain. We use the same PRO NER system, without training the domain specific models separately. Tables 4 and 5, respectively, show the performance of PRO NER in the two domains (where *P*, *R*, and *F* represent precision, recall, and F-measure, respectively). It is evident that PRO NER has achieved fairly high performance in both domains. This can validate to some extent the portability of our system.

Second, the results also show that our system performs slightly better in the cell phone domain in both the closed test and the open test. This is due to the fact that there are more challenging ambiguities in the digital domain owing to more complex product taxonomy and more flexible variants of PRO NEs.

Table 4 Results in digital domain ($\beta = 8$)

PRO NER	Closed test			Open test		
	P	R	F	P	R	F
PRO	0.864	0.799	0.830	0.762	0.744	0.753
TYP	0.903	0.906	0.905	0.828	0.944	0.882
BRA	0.824	0.702	0.758	0.723	0.705	0.714

Table 5 Results in cell phone domain (β = 8)

PRO NER	Closed test			Open test		
	P	R	F	P	R	F
PRO	0.917	0.935	0.926	0.799	0.856	0.827
TYP	0.959	0.976	0.967	0.842	0.886	0.864
BRA	0.911	0.741	0.818	0.893	0.701	0.785

5.3.3 The performance comparison between HHMM-1, HHMM-2, integrated model, and ME model on PRO NER

Xiong et al. (2004) used a two-layer ME model in organization name recognition, and obtained satisfactory performance. We use this method in PRO NER. The Maxent Toolkit (http://homepages.inf.ed.ac.uk/s0450736/maxent_tool-kit.html) is used in our experiment. Due to the nesting structure of the three kinds of PRO NEs, we train two ME models to, respectively, recognize BRA and TYP in the inner layer and PRO in the outer layer. The feature selection of the ME model is consistent with that of the integrated model. The feature types include the word forms and POS tags in the context (window size = 5 words), the NE tag (NE and PRO NE) of the last position, the surface feature of the word (CWF), brand list, etc.

We compared the integrated model and the ME model. Table 6 gives the experimental results, where "1", "2", and "1 + 2", respectively, denote HHMM-1, HHMM-2 and the integrated model; and "*P*", "*R*", and "*F*", respectively, denote precision, recall and F-measure. It is evident from Table 6 that all three HHMM models outperform the ME model according to the F-measure score. The reasons may be as follows. The ME model for PRO NER processes the input sequence from left to right. This sequential recognition mode may result in the accumulation of recognition errors. Information on different layers cannot complement each other. On the contrary, the HHMM-based PRO NER approach can integrate the constraint information in each layer and among different layers; as a result, it has a more powerful modeling ability for PRO NEs which have nested structures and variant composition and lengths. In addition, Table 6 shows that HHMM-1 tends to result in low precision and high recall, while HHMM-2 yields high precision and low recall; therefore, neither of them can perform well enough to attain a high F-measure. As discussed in Sect. 5.3.1, it is clear that the integrated model can benefit from

Table 6 Comparison between integrated model and ME model for PRO NER

	PRO			TYP			BRA		
	P	R	F	P	R	F	P	R	F
1	0.63	0.84	0.718	0.70	0.94	0.800	0.74	0.73	0.737
2	0.83	0.70	0.760	0.93	0.78	0.851	0.83	0.68	0.743
1 + 2	0.78	0.81	0.797	0.84	0.90	0.869	0.82	0.70	0.758
ME	0.81	0.59	0.683	0.82	0.43	0.564	0.58	0.62	0.60

combining the two individual models and attain better F-measures for all three kinds of PRO NEs.

6 Conclusions and future work

In this paper, we analyze the characteristics of PRO NEs and establish a specification for building PRO NE annotated corpora. Using the corpus we built, studies on automatic PRO NER methods in Chinese text are explored. Experimental evaluation of the proposed HHMM-based method indicates that this is a promising line of research. However, in order to make the HHMM-based method suitable for business-informatics applications, improvements need to be made in several areas.

First, the specifications on PRO NE annotation can be further improved to obtain more consistent annotation. Second, we can try to use long distance dependency information in PRO NER to remove some ambiguities that are difficult to remove in the current system. Third, the processes of segmentation, POS tagging, general NER and PRO NER can be integrated in order to alleviate error spreading.

Acknowledgments This work is supported by the National High Technology Development 863 Program of China under Grant No. 2006AA01Z144, the National Natural Science Foundation of China under Grant No. 60673042, and the Natural Science Foundation of Beijing under Grants No. 4052027 and 4073043. This research is also carried out as part of a cooperative project with Fujitsu R&D Center Co., Ltd. We would like to thank Dr. Hao YU, Dr. Yingju XIA, and Dr. Fumihito Nishino for helpful conversations and feedback on the corpus. We would like to thank Dr. Yang LIU of the University of Texas at Dallas, Dr. Ying ZHAO of Tsinghua University, and Mr. Matthew Trueman for their useful suggestions for modifying earlier drafts of the paper. We are grateful to the anonymous reviewers for very helpful comments on an earlier draft. Their insights and suggestions have led to many improvements in the paper.

Appendix: Peking University's TagSet for POS Tagging Chinese Texts (Yu et al. 2003)

Code	Chinese name	English name	Code	Chinese name	English name
a	形容词	Adjective	n	名词	Noun
b	区别词	Noun-modifier	o	拟声	Onomatopoeia
c	连词	Conjunction	p	介词	Preposition
d	副词	Adverb	q	量词	Measure word
e	叹词	Interjection	r	代词	Pronoun
f	方位词	Localizer	s	处所词	Place noun
g	语素	Morpheme	t	时间词	Temporal noun
h	前接成分	Head/Prefix	u	助词	Particle
i	成语	Idiom	v	动词	Verb
j	简称略语	Abbreviation	w	标点符号	Punctuation mark
k	后接成分	Tail/Suffix	x	非语素字	Non-morpheme character
l	习用语	Collocation	y	语气词	Modal/sentence-final particle
m	数词	Number	z	状态词	Stative adjective and adverb

References

Aberdeen, J., et al. (1995). MITRE: Description of the ALEMBIC system used for MUC-6. In *Proceedings of the Sixth Message Understanding Conference* (*MUC-6*) (pp. 141–155).

Bick, E. (2004). A named entity recognizer for Danish. In Lino et al. (Eds.), *Proceedings of the 4th International Conference on Language Resources and Evaluation* (*LREC2004*), Lisbon (pp. 305–308).

Bikel, D. M., Miller, S., Schwartz, R., & Weischedel, R. (1997). Nymble: A high-performance learning name-finder. In *Proceedings of the Fifth Conference on Applied Natural Language Processing* (pp. 194–201), ACL.

Borthwick, A. (1999). A maximum entropy approach to named entity recognition. PhD Dissertation. Computer Science Department, New York University.

Carletta, J. (1996). Assessing agreement on classification tasks: The Kappa statistic. *Computational Linguistics, 22*(2), 249–254.

Collier, N., Nobata, C., & Tsujii, J. (2000). Extracting the names of genes and gene products with a hidden Markov model. In *Proceedings of the 18th International Conference on Computational Linguistics* (COLING'2000), Saarbrucken, Germany (pp. 201–207).

Fine, S., Singer, Y., & Tishby, N. (1998). The hierarchical hidden Markov model: Analysis and applications. *Machine Learning, 32*(1), 41–62.

Jelinek, F., & Mercer, E. L. (1980). Interpolated estimation of Markov source parameters from sparse data. In D. Gelsema & L. Kanal (Eds.), *Pattern recognition in practice*. North-Holland.

McCallum, A., Freitag, D., & Pereira, F. (2000). Maximum entropy Markov models for information extraction and segmentation. In *Proceedings of the Seventeenth International Conference on Machine Learning* (ICML-2000), Stanford, CA (pp. 591–598).

Niu, C., Li, W., Ding, J., & Srihari, R. K. (2003). A bootstrapping approach to named entity classification using successive learners. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics* (ACL 2003) (Sapporo, pp. 335–342).

Pierre, J. M. (2002). Mining knowledge from text collections using automatically generated metadata. In *Proceedings of Fourth International Conference on Practical Aspects of Knowledge Management* (*PAKM2002*), Vienna (pp. 537–548).

Sekine, S., Grishman, R., & Shinou, H. (1998). A decision tree method for finding and classifying names in Japanese texts. In *Proceedings of the Sixth Workshop on Very Large Corpora*, Canada, http://cs.nyu.edu/~sekine/papers/wvlc98.pdf.

Sigel, S., & Castellan, N. J. (1988). *Non-parametric statistics for the behavioral sciences* (2nd ed.). McGraw-Hill.

Wu, Y., Zhao, J., & Xu, B. (2003). Chinese named entity recognition combining statistical model with human knowledge. In The Workshop attached with 41st ACL for Multilingual and Mix-language Named Entity Recognition: Combining Statistical and Symbolic Models, Sappora (pp. 65–72).

Xiong, D., Yu, H., & Liu, Q. (2004). Tagging complex NEs with Maxent models: Layered structures versus extended Tagset. In *Proceedings of the First International Joint Conference on Natural Language Processing* (IJCNLP-04), Sanya (pp. 638–643).

Yi, E., Lee, G. G., & Park, S.-J. (2004). SVM-based biological named entity recognition using minimum edit-distance feature boosted by virtual examples. In *Proceedings of the First International Joint Conference on Natural Language Processing* (IJCNLP-04), Sanya (pp. 22–24).

Yu, S., Duan, H., Zhu, X., Swen, B., & Chang, B. (2003). Word segmentation, POS tagging and phonetic notation. *International Journal of The Chinese and Oriental Languages Information Processing Society, 13*(2), 121–159.