International Journal of Knowledge and Language Processing
Volume 2, Number 4, October **2011**

www.ijklp.org
KLP International ©2011 ISSN 2191-2734
pp.1–27

Extracting Person Descriptions from Chinese Newswire through Ontological Semantic Calculation

Sujian Li[1], Wenjie Li[2]

[1] Key Laboratory of Computational Linguistics
Peking University
Haidian District, Beijing, China
lisujian@pku.edu.cn

[2] Department of Computing
The HongKong Polytechnic University
HongKong, China
cswjli@polyu.edu.hk

Received August 2011; revised October 2011

ABSTRACT. *It is a challenging task to find and gather the detailed descriptions of what we are interested in. We present work on use of patterns and ontological semantic knowledge to extract descriptions for person entities. Patterns are generalized through observation of annotated corpus. A mechanism of semantic calculation is given according to domain-specific corpus and HowNet. This paper illustrates that syntactic and semantic information should be combined to accomplish the task. The quantification of ontological semantic can provide further filtering with the guidance of patterns. At last we experimentally evaluate the proposed method, with promising results.*

Keywords: person description, information extraction, ontological semantic

1. **Introduction.** With the explosion of information, it is important to gather and digest detailed descriptions of named entities. With exact descriptions, people can quickly understand and grasp the content which they are concerned with. Among all kinds of named entities, person entities are always playing important roles within news reports since almost every piece of news are talking about a certain person and his/her related properties or affairs, people become more and more interested in tracking prior descriptions of a given person. In Document Understanding Conference 2004 [DUC 2004], the task has been proposed to create a short summary aiming at the question "Who is X" where X is the name of a person. In fact, this can be seen as the process of extracting and organizing a person's descriptions.

Extracting descriptions that relate to a person aims to automatically pull out relevant information from large volumes of texts and is similar to some tasks mentioned both in MUC (Nancy,1998) and ACE program [2000], which belong to the category of information extraction (IE). However, the IE systems usually pre-specifies various attributes and relations for entities, while the task mentioned in this paper is data-driven and descriptions are not defined in advance. For example,

Example 1: 中共中央政治局委员、中央书记处书记**温家宝**在贵州省委书记**刘方仁**等陪同 下 ,...

Example 2: 巴基斯坦穆斯林联盟（谢里夫派）候选人、原最高法院大法官 **穆罕默德 拉斐克 塔拉尔**，…，当选巴基斯坦第九任总统，…

Example 3: １９８８年４月，玉环县城关镇招聘土地管理员，**陈云峰**以优异的成绩被录用了。

Example 4: 今年２４岁的**颜旭**１９９４年３月加盟江苏三毛集团，现在是集团下属三星级 宾馆三毛大厦的一名员工。

Here we can see that descriptions are always phrases which are semantically relevant to a person and help to identify the person's title, occupation, age and so on.

Acquisition of descriptions is the foundation in many applications, and some state-of-the-art systems are using the finite state automata to carry out extraction task with linguistic rules either derived from training corpus or specified manually. [Radev 1999, Liu 2000, Hideo 2000, Joho 2001] mainly adopts heuristic rules and pattern matching techniques to extract entity description information, such as organization, place, person and so on. The methods can be easily implemented, but they are efficient in collecting an initial set of descriptive phrases, laying foundation for further processing. Of course it is also noted that the methods mentioned are weak in that mainly structural information is concerned while the semantic information is neglected.

For example, we cannot get precise descriptions by simply using heuristic rules. For example,

Example 5: **郝海东**，…，当选金球奖。(Golden Ball Award)

The sentence structure of this example is similar to example 2, but "金球奖" is a prize name which can't serve as a person description. For such situation, further improvement is expected for our description extraction task. Here, we propose to combine existing semantic resource and corpus statistics to measure the association between terms and person entities, called Term-Entity association, to verify those description candidates, which are extracted using heuristic rules. In example 5, the association between "金球奖" and a person entity is expected to be quantitatively small and excluded from the description set. Nevertheless, it is a tough job to evaluate semantic association between terms and person entities. Ontology and semantic networks may be used to support measurement of semantic association [Gaizauskas 1997, Guarino 1998]. However, it is laborious to produce the resources from scratch.

We propose to adopt HowNet, an ontological semantic network, as the basis to conduct calculation of Person-Entity association. HowNet reflects the semantic association between terms and entities, while the statistical information of term distribution surrounding person entities is considered. These two kinds of information are used to tune the Term-Entity association, which represents how likely a term tend to be the descriptions of person entities. In order to solve the problem that some terms are absent from the semantic knowledge base, an effective approximated method is proposed to improve the calculation of Term-Entity association. Based on the Term-Entity associations, the candidate descriptions are verified and ineligible ones are filtered out. Experiments show that the performance of description extraction is enhanced.

The rest of the paper is organized as follows. In section 2, we survey the related works on description extraction and corresponding techniques. In section 3 and 4, we delineate our system model of extracting person descriptions in news documents. We combine the methods of pattern matching and ontological semantic calculation. Then in section 5, how to conduct ontological semantic calculation will be illustrated. In section 6, the technique details of description extraction are introduced. Section 7 will present and evaluate the experimental results. And section 8 concludes the paper.

2. **Related Works.** Some related works have been conducted on descriptions. Entity descriptions are descriptive phrases which provide information about an entity. For English, Appositive and predicate nominative constructions are always entity descriptions. [Radev 1999] mainly adopts the simple pattern matching techniques to obtain descriptions and categorize them by "trigger terms". In [Liu 2000], Linguistic analysis is also employed to mine the descriptions of phrases/queries. Specifically, the work is based on simple patterns such as *is a*, *and other*, *such as*, *especially*, *including*, etc., in order to retrieve descriptions from free text documents. In addition to using linguistic patterns only, [Hideo 2000] combined ranking techniques to score the sentences, e.g., a number of common terms and the position of sentences found in the document are considered to rank the sentence. In [Aholen 1998], data mining methods based on generalized episodes and episode rules are applicable to text analysis tasks such as descriptive phrase extraction. Episode is a collection of feature vectors with a partial order. Weighting scheme is also introduced to help in pruning out redundant or non-descriptive phrases. This approach is now useful for Finnish, a language that has the relaxed order of words in a sentence. In addition, cooccurrence clustering and association rule mining algorithms have been used to learn phrase definitions in [Nguyen 2003]. [Schiffman, 2001] has focused on the use of the verbs for the specific purpose of finding strong subject-verb associations, with a view towards selecting a clause or sentence, for producing biographical summaries, and semantic information from WordNet is combined to prune and merge only appositive descriptions.

The systems introduced above are designed to produce structurally simple fragment

parses. These simple fragments can be described with finite-state grammars that can be processed easily, robustly and quickly. However, some aspects about semantics need to be noted and used to improve the systems further. In IE systems, traditional semantic analysis in NLP is usually limited to finding the predicate argument structure of a small set of core propositions, still with a bad performance [Douglas 1999]. To efficiently process large volumes of real-world texts, semantic knowledge is borrowed by building a domain specific ontology. In computer science, an ontology is a formal, explicit specification of a shared conceptualization according to Gruber (1993), which can guide the tasks by restricting them to a specific domain. And unlike "rigid templates", ontology can provide them with knowledge inference and conceptual relationship. In [Nirenburg 2003], ontological knowledge is used in formulating lexicon and fact repository entries, and helps the referring expressions link to their corresponding entities. In [Luke 1996], a set of simple HTML ontology extensions can provide semantic organization to help in gleaning knowledge from the Web. In [Eivind 2003], a financial ontology has been made to state different concepts and relations, which can be queried and used by the agent as a guide to know what financial information to extract. [Wu 2003] has designed a domain ontology for analyzing and gathering semantic information of a class of articles, and the performance of ontology-based method is much better than random-selection method. Ontology is not the solution, but just part of the solution. Ontology-based methods are mainly divided into two types. One is to construct ontology through (semi-) automatic methods or manually [Wu 2003]; the other type relies on machine learning and automated language techniques to extract concepts and ontological relations [Roberto 2003]. Although various techniques of automatically constructing ontology have been reported and tested, there is still a long way to go. It's also difficult to construct manually from scratch. In addition, better no ontology than ontology without a good theory. Thus we expect to make use of existing ontological framework to provide semantic knowledge that is tailored to the news corpora.

HowNet developed by Dong et al (Dong and Dong 1999) is the best publicly available resource on Chinese semantics. It separates the concept network with the lexicon and every entry in the lexicon is explicated by several concepts. Based on HowNet, some researchers have conducted semantic computation to measure the relations between two words [Liu 2002, Li 2002]. This enlightens us to conduct semantic computation based on task specific ontology and corpora, which provide guidance to obtaining entity descriptions.

3. **Ontological semantic.**

3.1. **Introduction to Ontological Semantic.** An ontology is generally regarded as a designed approach to describe entities and their semantic relationship in some domain of interest (Guarino, 1998), which is used to deal with classes directly mapped to set of entities (instances) in some universe of discourse. Ontology itself is language independent. It is composed of concepts and relations within a specific domain. The left

part in figure 1 shows a portion of ontology structure surrounding the concept PERSON. A square node represents a concept and an arrow represents a relation.

However, Ontology can't extend beyond language barriers, and it needs a mapping from concepts to some concrete linguistic units. Semantics is a subfield of linguistics that is traditionally defined as the study of meaning of words, phrases and other linguistic units. Some perspective holds that the meaning of words can be analyzed by the study of relations between different linguistic expressions, which can be seen as a semantic network. Semantic networks are a common type of machine-readable dictionary. An example of a semantic network is WordNet [Miller 1995], which organized content words into sets of synonyms. Other relations between terms include hyponymy/meronymy hierarchies and don't put much emphasis on other relations. The middle part of figure 1 represents a sample of a semantic network, in which the oral nodes represent terms and the arrows represent relations of terms.

According to Maedche (2003), there should exist a structure <O, L> where O is an ontology structure and L means a corresponding lexicon. To build such a structure, concepts, relations and the mapping from terms to concepts and relations must all be considered in advance and it is very difficult to build from scratch. On one hand, we make use of the ontological organization. On the other hand, we adopt semantic meanings of terms in a semantic network which can be defined by ontological concepts. As in the right of figure 1, oral nodes which represent terms are linked by various relations, and the whole network has embodied the relations of terms which in fact are instantiated by certain concepts. Thus, with domain concepts and their relations as framework, we can acquire relations between terms.

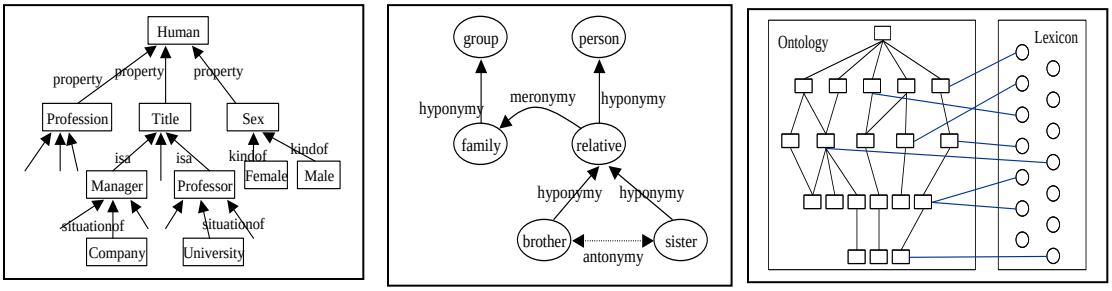

FIGURE 1. sample of ontology, semantic network, and ontological semantic

We find that HowNet is a good resource to represent ontological semantic relations of terms. Hownet has provided a semantic lexicon now available, which has illustrated inter-conceptual relations and inter-attribute relations of concepts. For our task of extracting descriptions, we note that those concepts which have close relationship with the person concept are always concepts which will be instantiated terms acting as person descriptions. However, so many relations exist between terms that it is difficult to qualitatively define them. Thus, we hope to construct a computational mechanism to

measure the relations of terms according to HowNet, through which we can obtain the descriptions. In section 4 we will describe in detail the calculation methods.

3.1. **HowNet.** HowNet is a Chinese thesaurus, containing inter-concept relations and inter-attribute relations among concepts [Dong 2000]. Each term in HowNet represents a concept and is defined by various related sememes, which are the smallest semantic units that cannot be further partitioned. For example, "human|人" is a sememe, which has its hypernym sememe "AnimalHuman|动物" and is described by other sememes "name|姓名, wisdom|智慧, …". "经理(manager)" is a term that is defined by a set of sememes "human|人,#occupation|职位,official|官,commercial|商". The terms, sememes and their descriptive sememes are illustrated in figure 2. In such a way, HowNet actually provides us a semantic network of sememes, from which the parameters, such as *closeness*, *relatedness*, and *relevancy* between sememes, could be computed step by step and the term relevancy H_Rele could be obtained in turn.

Definition 3. ***Closeness*** (indicated by sim) between the two sememes, s_1 and s_2, is to calculate how close s_1 is hierarchically related to s_2 if s_1 and s_2 belong to the same category.

All sememes in HowNet are categorized into 7 categories, including event, entity, antonym, attribute (value), converse, quantity (value) and secondary feature, which can be further sub-categorized. Sememes in a category are organized in a tree structure. Except for the hierarchical relations in one tree, two sememes have other associations which can be measured by *relatedness*

Definition 4. R***elatedness*** (indicated by $asso$) between the two sememes, s_1 and s_2, is to calculate how they relate to each other according to their descriptive sememes and hypernyms/hyponyms. For example, there is *relatedness* between "human|人" and "age|年龄" although they are in different categories.

Definition 5. R***elevancy*** (indicated by $rele$) between sememes s_1 and s_2 is to calculate the association of the two sememes according to their *closeness* and *relatedness*.

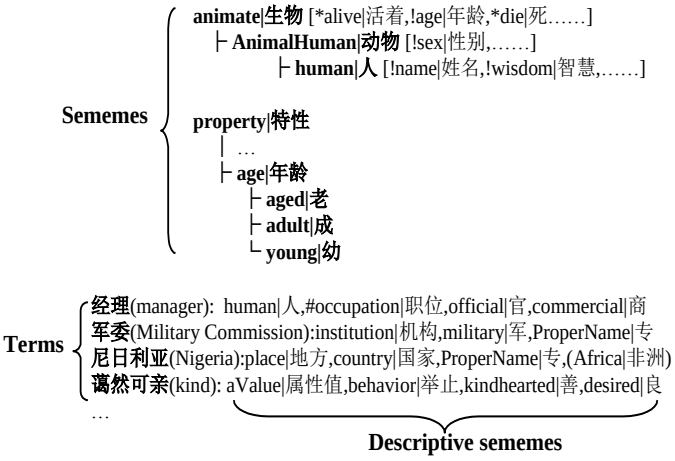

FIGURE 2. *relatedness* Examples of sememes, term s, and the closeness, relatedness, relevancy between them

4. **Model of Person Description Extraction.** In this section, we first clarify several notions used in the process of extracting entity descriptions. Then the system architecture is shown how to extract person descriptions.

4.1. **Some Notions.** Each system has its own definition of what is a person description. In **[Radev 1998]**, the function "D = DescriptionOf (E)" defines the relation to be the one between a named entity E and a noun phrase D, clearly identifying the named entity. The noun phrases are always pre-modifiers and appositions, ranging from simple nouns to longer expressions, even extended to the scope of those appearing in relative clauses. Rather than determining in advance what sort of information belongs to descriptions, our approaches are data-driven. In the text, the presence of some properties or events in the immediate context of a proper name can be used to provide confirming or criterial evidence for identifying a name [David 1993]. Thus, **Person Description** is a linguistic unit in the context surrounding a person name, used to describe a person entity, such as its properties and relations with other person entities.

A linguistic unit means that a description must represent a complete meaning. That is, a description can be a word, a phrase or a relative clause, which have been preprocessed through the techniques of word segmentation and shallow parsing. Rather than predefining what attributes should be extracted, description extraction depends on how persons are described in news reports. Mainly the characteristics we focus on include title, age, profession, nationality and et al, which can be used to differentiate among entities. In addition, a description is time-relevant and has its own occurring time. Without time characteristic, a description is inaccurate and sometimes even meaningless. For the same description of a person, there are often different expressions, which should be uniformed. Table 1 illustrates some examples of descriptions produced in our research.

TABLE 1. Examples of descriptions

Example			**Characteristic**	**Category**
Person	**Description**	**Time**		
卡翁达 (Kaunda)	赞比亚前总统 (Zambian former president)	2/6/1998	Title	Properties
马罗尼 (Maroni)	今年23岁 (23 years of age this year)	3/6/1998	Age	
安杰依 (Andrzej)	波兰人 (a Polish)	2/6/1998	Nationality	
孙茂庆 (Sun maoqing)	新华社记者(reporter of Xinhua News Agency)	2/6/1998	Profession	
武雪芹 (Wu xueqin)	炮团随军家属(family dependent in the military troop)	3/6/1998	Relative	Relations

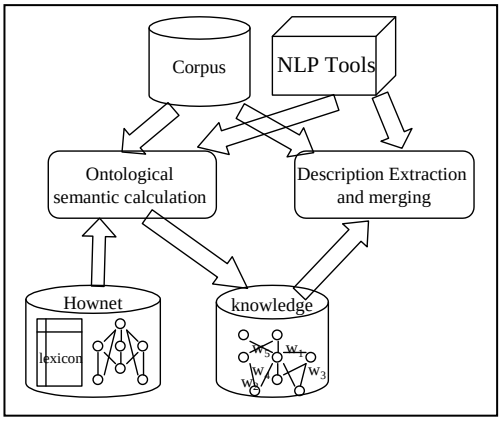

FIGURE 3. Architecture of extracting descriptions

A description is composed of terms and function words. In text a ***term*** is a semantic concept with definitions in HowNet lexicon. For example, in example 1, "president", "age" are both terms defined in HowNet, which are also constituent parts of descriptions. In order to extract person descriptions, we need to evaluate the association between terms and person entities. Here we give definitions as follows:

Definition 1. ***Person Entities*** are instantiated from the *PERSON* concept in semantic knowledge, and has the form of person names in text. For example, **Kaunda, Maroni**, and **Andrzej** are all called person entities.

Definition 2. ***Term-Entity Association*** represents the association between terms and entities. Here, we focus on the person entities.

In this paper we restrict our research of extracting descriptions to two aspects: (1) we use patterns to obtain possible linguistic units as candidate descriptions. Patterns refer to the structural relationship that descriptions bear to person entities. (2) As for how to identify whether a linguistic unit characterizes a person, we need to introduce semantic knowledge.

4.2. **Architecture Overview.** Figure 3 illustrates the system architecture used for extracting and processing descriptions. The first concerns the process of extracting descriptions. On the one hand, it will make use of the common techniques of natural language processing, which include word segmentation, POS tagging, chunk parsing. With the results of those tools, we can acquire some extraction patterns to give the scope of locating where descriptions are. On the other hand, we need to utilize the term relations to further determine what appropriate descriptions are.

The second key area is the construction of ontological semantic calculation, through which relations can be quantitatively measured. Here we suppose that those terms having close relations with the terms instantiated from the person concept are adaptable to act as person descriptions. Then, every relation is given a numerical value of relevancy which measures the degree of relatedness. Relevancy is not only restricted to representing the similarity of the term that belong to the same category, but also measuring the relation of

terms that belong to different category. For example, the terms of “nurse” and “doctor” have high similarity and relevancy, while “hospital” and “doctor” has no similarity but have high relevancy. Here we can see that relevancy is a more appropriate means to acquire descriptions. In addition, our goal targets for newswire, and term usage in news corpus is also a consideration for constructing the quantitative knowledge base.

The final concern is the post-processing of descriptions. Descriptions should have a uniform format which can be referenced in later use. Because descriptions are acquired from multiple documents, there inevitably exist the phenomena of the repeated descriptions for the same person, or similar descriptions with the same reference for the same person. For such redundancy existing, we adopt a simple strategy of merging. At the same time, time information is attached as a feature of a description for easy tracing later. These will be introduced in detail in section 5.

5. **Ontological semantic Calculation.** Then ontological semantic knowledge is needed for further improvement. Due to the existing inadequacies of adopting an ontology, we propose an alternative mechanism of semantic calculation.

5.1. **Ontological Semantic Computation.** HowNet is used to conduct semantic computation of terms. Firstly, it has given a semantic network of sememes, and each sememe represents a certain concept. There is also a lexicon in HowNet, and every term is explicated by sememes. Based on the quantitative calculation of relations between sememes, we can get quantitative results of term relations, furthermore obtain the numerical values to quantify the relation between each term and person name in text. This process reflects how to measure for the terms in a person relevant domain. In addition, the news text has its own literal characteristics. Thus we collect news documents, which can provide us statistical data about term distribution surrounding person names in the news domain. Then HowNet as person-relevant domain knowledge and corpus as concrete language experience are a good combination to measure if a term should be selected as a person description.

Thus the whole flowchart is shown in figure 4. Firstly, we do syntactic processing for the corpus document, such as word segmentation, pos tagging etc. SRW calculation is done on the annotated corpus to illustrate the distribution of every term. Person names in texts are all uniformed as the person concept in HowNet. And SemReve calculation is conducted to give a measure for every relation between a term and the person concept based on HowNet. Then, according to the results of SRW calculation and SemReve calculation, we can further tune the relevancy between a term and the person concept. It is re-estimated the contribution of every sememe to becoming a description and all terms are recalculated to get new numerical values of relevancy with the person concept. Thus, a mechanism of person relevancy calculation is built to embody the relations of person relevant things/events in the news domain. And these terms with person relevancy constitute the quantitative knowledge, which is used for extracting descriptions. How to use the knowledge will be introduced in next section.

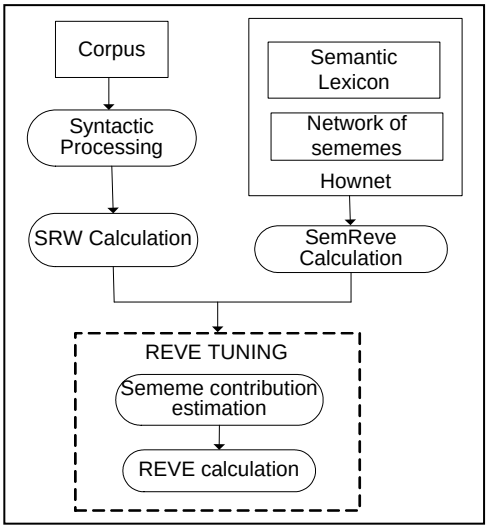

FIGURE 4. flowchart of person relevancy calculation

5.1. **SRW Calculation of Terms.** Descriptions must have some literal characteristics in the text, which can be embodied in the corpus. Documents in different domains have different expressive styles to characterize their persons. Firstly, the text in one news document is relatively canonical so that every entity is often given a clear characterization once it occurs the first time. Secondly, a news article is always brief, and descriptions tend to concentrate surrounding the corresponding entities rather than disperse over the whole document. Lastly, for news documents a person's characterization mostly expand with person names as anchors.

According to these characteristics of newswire, we suppose that descriptions always occur near person entities. That is, if a word often occurs near a person name, there is more possibility for it to belong to a description. And those which distribute evenly in the corpus are unfit for being descriptions. A corpus has been segmented and named entities have also been tagged correctly, and then we call this corpus as the original corpus. From the original corpus we extract the sentences which include person names, and take them as a reference corpus. SRW calculation makes use of the results of syntactic processing to give the distributions of person relevant things/entities in the news corpus. Under the hypothesis above, those words which can be used as descriptions should have a higher distribution density in the reference corpus than in the original corpus. Then, the frequencies of every word in both corpora are recorded and significance ratio of a word (SRW) can be calculated out. The formula of SRW is given as follows,

$$SRW(w_i) = \frac{density\ of\ w_i\ in\ reference\ corpus}{density\ of\ w_i\ in\ originial\ corpus} (1 \le i \le n) = \frac{occurring\ number\ of\ w_i\ in\ reference\ corpus\ /\ N}{occurring\ number\ of\ w_i\ in\ original\ corpus\ /\ M} \quad (1)$$

Where w_i represents a word, and there are totally n distinct words in the lexicon. There are totally N and M words respectively in the reference corpus and the original corpus. The greater the SRW of a word w_i, more probably it can be a description.

5.2. **SemRele Calculation of Terms.** SRW gives the tendency of terms as descriptions in the corpus, while HowNet provides the resource to calculate semantic relevancy of all terms. This kind of semantic relevancy conforms to the conceptual structure in our mind. More semantically relevant a term is to a person name, more possible the term would be a description. Thus a mechanism of semantic relevancy computation is useful for extracting descriptions. HowNet has provided a set of sememes and a Chinese lexicon, in which every term is explained by a set of sememes. The whole structure can be simply formalized as follows:

$$\begin{aligned} &SS = \{s_1, s_2, \ldots, s_m\} \\ &T = \{t_1, t_2, \cdots, t_n\} \\ &REL = \{*, @, ?, !, \sim, \#, \$, \%, \wedge, \&, NULL\} \\ &def(t_i): r_{i1}s_{i1}, r_{i2}s_{i2}, \cdots, r_{ik}s_{ik},\ r_{it} \in REL,\ s_{ij} \in SS (1 \le i \le n; 1 \le j, t \le k) \end{aligned} \tag{2}$$

where *SS* represents the set of the sememes which includes *m* elements; *T* represents the set of terms in *HowNet* whose size is *n*; *REL* is the set which describe relations between a concept and a sememe or relations between sememes. For every term t_i, its definition $def(t_i)$ is composed by *k* items, and each item includes a relation symbol in *REL* and a sememe in *SS*.

A sememe is a basic semantic unit that is indivisible in HowNet. About 1,667 sememes are extracted to compose an elementary set which is the basis of the Chinese glossary, as over 100 kinds of chemical elements constitute all the substances in nature. All sememes construct a network structure. Figure 5 shows a sample structure of sememes. Every sememe (boldfaced) can also have its explicative sememes in square brackets.

animate|生物 [*alive|活着,!age|年龄,*die|死,*metabolize|代谢]
　├ **AnimalHuman|动物** [!sex|性别,*AlterLocation|变空间位置,*StateMental|精神状态]
　　├ **human|人** [!name|姓名,!wisdom|智慧,!ability|能力,!occupation|职位,*act|行动]

property|特性
　| …
　├ age|年龄
　　├ aged|老

FIGURE 5. Sample structure of sememes

The definition of a term is composed of sememes. Here examples were given to show how a term is represented by sememes. Terms are located before the punctuation ":", and their definitional sememes are located after ":".

Thus we can see that the semantic relevancy calculation of terms is based on the semantic calculation of sememes. If two sememes belong to the same category and have hypernymy or hyponymy relations, we can give a numerical value called similarity to quantify their relation. If two sememes belong to different category, their relation can also be quantified by associativity, through calculating the similarity among their

explicative sememes, their hypernyms|hyponyms or explicative sememes, and hypernyms|hyponyms of explicative sememes.

蔼然可亲(kind): aValue|属性值,behavior|举止,kindhearted|善,desired|良
经理(manager): human|人,#occupation|职位,official|官,commercial|商
军委(Military Commission): institution|机构,military|军,ProperName|专,(China|中国)
尼日利亚(Nigeria): place|地方,country|国家,ProperName|专,(Africa|非洲)
…

FIGURE 6. Examples of term definition

Liu[2002] ignored the network relations of the ontology and thus the result cannot describe the relevant relations precisely. Li[2002] adopted two different architecture in one system and didn't normalize the result, which presented some inconsistency. We overcome those deficiencies and firstly compute the similarity between any two sememes s_1 and s_2 as formula 3.

$$sim(s_1, s_2) = \begin{cases} \alpha/(dist(s_1, s_2) + \alpha) & tree(s_1) = tree(s_2) \\ 0, & tree(s_1) \neq tree(s_2) \end{cases} \quad (3)$$
$$s_1, s_2 \in SS'$$

where, for any two sememes s_1,s_2 in the sememe set *SS*, $sim(s_1,s_2)$ represents similarity between s_1 and s_2. $tree(s_1)=tree(s_2)$ represents that the two sememes s_1 and s_2 are in one tree structure and their similarity is inversely proportional to their distance. During calculation we take the parameter of α as 0.5.

In order to compute associativities of two sememes, we need to expand the current sememe in two directions. One is to expand to the hypernyms of a sememe's explantory sememes which is called Horizontal Associative Expansion (*HAE*), the other expansion is to the explantory sememes of the hypernyms which is called Verticle Associative Expansion (*VAE*). We compute associativities between sememes s_1 and s_2 as formula (4).

$$\begin{cases} asso(s_1, s_2) = \beta_1 avg(\sum_{s_i \in ext(s_1)} sim(s_i, s_2)) + \beta_2 avg(\sum_{s_j \in ext(s_2)} sim(s_1, s_j)) \\ ext(s_j) = \{s_i \mid s_i \in HAE(s_j) \cup VAE(s_j)\} \\ avg(\sum_{s_i \in ext(s_1)} sim(s_i, s_2)) = \sum_{s_i \in ext(s_1)} sim(s_i, s_2) / |ext(s_1)| \\ \beta_1 + \beta_2 = 1 \end{cases} \quad (4)$$

Where $ext(s_j)$ is sememe s_j's expansion set in which every sememe s_i belongs to *HAE* and *VAE* and $|ext(s_j)|$ means the total number of sememes in this set. We compute two average numbers, one is that of similarity between s_1 and every sememe in $ext(s_2)$, and the other is that of similarity between s_2 and every sememe in $ext(s_1)$. β_1 and β_2 are normalized parameters with which the associativity $asso(s_1,s_2)$ will be between 0 and 1. In addition, the associativity of two sememes is supposed to have the symmetry characteristic, and we set the values of β_1 and β_2 both as 0.5.

The goal of computing the similarity and associativity between sememes is to get the

relevancy of terms. We suppose that semantic relevancy of two terms completely relies on the sememes. Formula 5 gives the calculation equation.

$$\begin{cases} Sem\,\mathrm{Re}ve(t_1,t_2) = \mathrm{Re}le\,(def\,(t_1), def\,(t_2)) \\ \mathrm{Re}le\,(def\,(t_1), def\,(t_2)) \approx \sum_{s_i \in def\,(t_1)} \max_{s_j \in def\,(t_2)} \mathrm{Re}le\,(s_i, s_j) \\ def\,(c) = \{s_i \mid REL(c, s_i)\} \\ \mathrm{Re}le\,(s_i, s_j) = w_s sim\,(s_i, s_j) + w_a asso\,(s_i, s_j) \\ w_s + w_a = 1 \end{cases} \tag{5}$$

Where *SemReve(t_1,t_2)* is the relevancy between two terms t_1 and t_2. *def(.)* is a set of explanatory sememes for a term. Then semantic relevancy calculation of terms will be converted to finding the highest relevant sememe pairs *Rele(s_i,s_j)* and sum their relevancy up. w_s and w_a are the weights of similarity and associativity between sememes respectively. With $w_s + w_a = 1$(here we take the experience value 0.8 and 0.2 respectively for w_s and w_a), we can get a relevancy *Rele(s_i,s_j)* whose value is between 0 and 1.

To calculate the relevancy of a term and a person name in the text, we use the semantic definition "human|人" to represent the person entity. Then according to formula 5, we can get the semantic relevancy value of every term with the person name.

5.3. **Tuning of Relevancy.** For every term in the corpus, we can calculate *SRW* and semantic relevancy (*SemRele*) with a person entity. Both *SRW* and *SemRele* of a term can reflect the tendency of being a description. There are two situations which needs improvement in judging whether a term belongs to a description. One is, that some terms exist in the corpus, but aren't defined in the lexicon. For example, "爱孙(loved grandson)" is not defined in the semantic lexicon, and its *SemRele* will be 0. We know that the term should in fact have a higher *SemRele*.

The other situation is data sparseness. That is some terms that never occurred in the training corpus, but they may perhaps have a higher *SemRele* according to the semantic definition. Then the *SRW* value with 0 will affect its selection as a description.

The reason of the former situation is that words in HowNet lexicon and in segmentation lexicon aren't completely the same. We need knowledge extension and add those new terms to the semantic lexicon of HowNet. That is, a semantic definition must be given to each term. Here, we adopt a simple strategy to implement this. If a new term is a proper name, according to the result of named entity recognition we give it an definition. For example, if the term is a place name, we will define it as "place|地方,ProperName|专". If a new term is not a proper name, firstly we will use backward maximum match method to segment it according to HowNet lexicon. Then, we take the definition of its last word as its definition. For example, the new term "爱孙(loved grandson)" will inherit the semantic definition of "孙(grandson)" and be assigned the semantic definition "human|人,family|家,male|男".

Term distribution can finally affect the distribution of sememes because terms are defined by sememes. Then we hope to combine term distribution with semantic definition and estimate the contribution of all sememes. For terms never occurring in

the training corpus, we can obtain its relevancy with person concept according to the definition of the term. And it will be up to the sememes in the definition whether a term is a description. From the training corpus, if the terms defined with some sememe can always be regarded as a description, the sememe will have a high contribution to being a description. Thus, it is concluded that the definitions of other terms can indirectly affect whether a term will be a description. For the latter situation, if the contribution of every sememe has been accurately given, we can ignore the SRW and judge only from the contribution of sememes whether the term is a description.

Then we need to estimate the contribution of every sememe and calculate the score of being a description for every term.

Firstly, Both SRW and semantic relevancy contribute to being a description for a term, and an initial score can be assigned according to formula (6).

$$\begin{aligned} &score(t_j) = aSRW(t_j) + bSem\,\mathrm{Re}le(t_j)(1 \le j \le n) \\ &a + b = 1 \end{aligned} \tag{6}$$

Where t_j means a term, $score(t_j)$ is an initial score of evaluating the relevancy between term t_j and a person entity, $SRW(t_j)$ is the SRW value of term t_j, and $SemRele(t_j)$ is the semantic relevancy of t_j. a and b represent what those two factors contribute to being a description. We suppose that SRW and semantic relevancy have the same contribution to becoming a description, then a and b are both set 0.5. That is, we score the relevancy of every term according to the average of SRW and SemRele. Thus, the term distribution can help to leverage the contribution of sememes.

Secondly, for every sememe, we collect all the terms whose definition consists of this sememe. The score of those terms will be summed up and averaged as the contribution of the sememe according to formula (7).

$$cont(s_i) = avg(\sum_{t_j \in Ts.t.s \in def(t_j)} score(t_j))(1 \le i \le m; 1 \le j \le n) \tag{7}$$

Where s_i means a sememe, $cont(s_i)$ is the contribution of sememe s_i to being a description. $def(t_j)$ is the definition of term t_j. Whenever the definition $def(t_j)$ includes sememe s_i, we sum $score(t_j)$ up. Then we get the average of the scores as the contribution of the sememe.

Thirdly, as formula (8), we can calculate the relevancy of every term with a person entity according to the contribution of sememes. At the same time, we can give a more reasonable value to evaluate the relevancy of those terms which have never occurred in the training corpus.

$$\mathrm{Re}ve(t_j) = avg(\sum_{s_i \in def(t_j)} cont(s_i))(1 \le i \le m; 1 \le j \le n) \tag{8}$$

where t_j means a term in the lexicon, $Reve(t_j)$ is the relevancy of term t_j which is used to measure how relevant the term is with a possible person occurring in the text. The relevancy is calculated by averaging all the contribution of those sememes which belong to the definition of the term t_j.

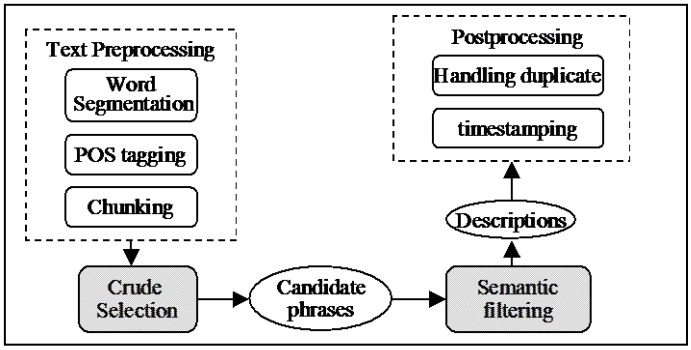

FIGURE 7: Flowchart of Extracting Descriptions

6. Details of Description Extraction. Figure 7 illustrates the whole pipelined processing to obtain descriptions. Description extraction should belong to the category of IE. As [Douglas 1999] proposed, there should be some core elements for every information extraction system. Then the first step is that news documents are preprocessed by all kinds of NLP tools, which can identify parts of speech, all kinds of chunks, and person names in text. These preprocessing techniques are an important prerequisite, which provides a fairly robust mechanism for producing text representation that can be effectively used for extracting entities and their descriptions. Text preprocessing will not be discussed in detail due to limited space.

Once the person names (a person name indicates a person entity) in the texts are identified, we should focus on occurrence of person entities, ignore unwanted text, and locate possible scope of descriptions. We notice that in most cases, those phrases, which give out characteristics of person entities, normally occur near the person names in the text, mostly in the same sentence. After sentence segmentation, we select those sentences that include person names for further consideration for extracting descriptions. This step can significantly reduce the data size and computation time without losing much useful information. In addition, even in one sentence, we should analyze syntactic structures surrounding person names and only pick out those phrases which occur before or after person names and often act as descriptions. According to the art of state survey, pattern matching can be an efficient means of locating possible descriptive phrases surrounding person entities from sentences with person names. With sentence selection and pattern matching, we conduct a crude selection to obtain candidate descriptive phrases.

Then, a set of candidate descriptions is obtained mainly through the analysis of syntactic structures. Furthermore, it is important to catch semantic meaning behind those patterns. A mechanism of relevancy calculation has just provided such functions. Each candidate is sent to conduct calculation and given a quantified representation which evaluates how much degree it achieves in being a eligible person description. Then those that don't pass through a certain threshold are filtered out.

Descriptions are extracted from a large number of news documents. Because one person entity may occur many times, its descriptions in different places sometimes are completely the same, sometimes literally similar with the same reference. We should consider how to handle the redundant information. In addition, we have known that time

is an important feature of descriptions. During description extraction, we also extract time information in the context to limit descriptions. Every description is placed a timestamp. With these postprocessing, each person entity has consistent descriptions with time features, which are very useful to the concrete application such as summarization task. For example, if we summarize for one document and can't find the detailed description for some person, then we can refer to its descriptions in certain time phrase for a reasonable explanation.

6.1. **Crude selection.** In the process of crude selection, the first thing to do is to locate the descriptive scope, which means possible scope of containing descriptions. With sentence segmentation, we can extract those sentences which include person names as a rough range. Sometimes, we need further locating, especially when there are several person names in one sentence. If these person entities have the juxtaposition relation, we take them as a whole and the sentence containing them is also the range of obtaining descriptions. Otherwise, segmentation words will be found which occur between two person names, and thus the sentence will be divided into several parts which give its own descriptive scope. For example,

Example 1:

泰(Thailand)/j 总理(Prime minister)/n 川·立派(Chuan Leekpai)/nr … 接见(have an audience with)/v … 中国贸促会(China Council for the Promotion of International Trade)/j 会长(chairman)/n 俞晓松(Yu xiaosong)/nr 为(as)/v 团长(head)/n 的(of)/u 中国(Chinese)/ns 经贸(economic & trade)/j 代表团(delegation)/n 。/w

In example 1, the sentence has been segmented and POS tagged(POS tag set can be referred in the appendix), and here the verb "接见(have an audience with)" will be selected as a segmentation point. The part before this point will be descriptive scope of the person name "川·立派(Chuan Leekpai)". The latter part will then be the descriptive scope of the person name "俞晓松(Yu xiaosong)".

For each person entity, we need further finer processing on descriptive scope. Pattern matching is widely used in extraction tasks with satisfactory results [Radev , Hu 2004]. According to [Joho 2001], this simple approach is much faster than one that requires complex operations such as parsing, and another reason is that it provides a means of determining lexical relations from corpora that are worthy of further exploration. A crucial step is generation of patterns. For English, description phrases (DP) are always extracted for entities (tagged as Entity) according to the following patterns.

…Entity, (a|the) DP, ….

…Entity is a DP

…Entity (is|was|are|were) (a|an|the) DP.

…DP, such as Entity …

So far, we haven't seen any patterns generalized to extract person descriptions for Chinese documents. Here we make statistics of the surrounding contexts of person

entities and the contexts are composed of syntactic tags such as POS tag, chunk category and so on. For one thing, we extract the maximal chunks surrounding a person name as the contexts. The maximal chunk means that the chunk doesn't be contained by any other chunks. Next, if a word doesn't belong to any existing chunk, then we take its POS tag as the context. According to those statistics, we generalize rules manually. For example,

Example 2:

"[忠诚(loyal)/a 的/u [共产主义(communistic)/n 战士(soldier)/n]NP]NP … 刘澜涛(Liu Lantao)/nr 同志(comrade)/n …".

Example 2 is a Chinese sentence with POS tags and chunking tags. Chunking tags will also be illustrated in appendix. Here the person name is "刘澜涛(Liu Lantao)" represented with "PERSON". Then the context of the person entity will be "NP … PERSON n …".

Table 2 shows some pattern examples, according to which we conduct pattern matching for context of person entities.

TABLE 2. Patterns for extracting descriptions

No.	Rule
1	{(NP|NZ|NT|QP|m|b|n|ns|nt|nz|vn|f|j|a|兼|驻|的)+ }DP PERSON
2	PERSON (等) {(NP|n|a|m|b|f)+}DP
3	PERSON ？是|当选| 作为|提名|评为|任|成为 ？{NP+}DP
4	{PERSON 的 n}DP PERSON
5	叫 PERSON 的 {n|NP}DP
6	{NP}DP －－ PERSON
7	作为 {NP}DP ， PERSON
8	像 PERSON (这样|那样) 的 {NP}DP
9	PERSON {QP|TP}DP
10	PERSON (－－) {NP}DP

Here PERSON represents a person entity occurring in the text. { }DP means that the bracketed content is a possible descriptive phrase. The symbol "( )+" means that the content in the brackets can occur one or more times and without plus mark "+" the content in the bracket is optional. Symbol "?" means that any word can fill there.

Through the patterns, descriptive scope can be further contracted. Those patterns can be seen as finite state automaton. And the output results through pattern matching were seen as candidate descriptions. There also exist some coreference phenomena in the corpus, and some descriptions distribute dispersedly. These dispersed rules are very difficult to generalize and ignored in this paper.

6.2. **Semantic filtering**. After crude selection, we have obtained a set of candidate descriptions. These phrases extracted only by patterns sometimes aren't eligible descriptions, which will decrease the system's precision. For example, if we extract descriptions from the clause "…宋庆龄 (Song Qingling)/nr 奖学金(Scholarship)/n…", according to rule (2) "奖学金" is a noun and should be extracted. However, this is not what we expect. We need further verification to see whether the results through pattern matching are reasonable. Weighted relations are used to measure the relevancy of candidate descriptions with its described person entity. Those with low relevancy will be filtered out.

In section 5 a mechanism of relevancy calculation has been introduced to measure the relation of terms with person entities. However, pattern matching will give some phrases occurring before or after person names as candidate descriptions. These phrases have been segmented, POS tagged and chunked. That is, in every candidate descriptive phrase, every word has its POS tag and is known which chunk it belongs to. According to the relevancy of terms and results of syntactic parsing, we need consider how to measure whether a candidate phrase is suitable for being a description. In a phrase, every chunk has its own head word which can represent the central meaning. Then we find head words for every chunk. Here for convenience, we define a simple method to obtain head words. For a chunk with a single word, it itself is taken as the head word. Other chunks we are concerned include noun chunks, verb chunks, and quantitative chunks. These three types of chunks play a major role in determining whether a phrase is a description. Their head words can be obtained iteratively as formula (9).

$$\begin{aligned} &\textit{Headword(Noun chunk) = Headword (the last chunk)} \\ &\text{Headword(Verb Chunk) = Headword(the first chunk)} \\ &\text{Headword(Quantitative chunk)= word with POS quantifier} \end{aligned} \tag{9}$$

Then, we can calculate the relevancy of every chunk with a person entity.

$$\begin{aligned} &relevancy(c_i) = relevancy(headword(c_i)) \\ &relevancy(cp) = \max_{c_i \in \{nc, vc, qc\}} (relevancy(c_i)) \end{aligned} \tag{10}$$

Then for a candidate phrase *cp*, every chunk c_i (noun chunk or verb chunk or quantitative chunk) in it has a relevancy value. We compare the relevancy values of all chunks and select the maximum one as the relevancy of the phrase. Then, we get the relevancy value of phrase *cp*, which is used to compare with the threshold (with experience value 0.2 here). If the value is greater than the threshold, it will pass through the semantic verification. Otherwise, the phrase will be deleted from the description set.

6.3. **Postprocessing of descriptions.** Every person appearing in documents have multiple descriptions associated with it. Merging is conducted to make descriptions consistent. Those repetitive descriptions for the same person entity are excluded. At the same time, every person has different descriptions at different time. During the merging, we will add the time as the property of the descriptions. In addition, we don't expect to solve the coreference problem. However, we will try to merge the descriptions for different names referring to the same person, and this will somewhat improve the performance.

6.3.1. **Duplicate Handling.** The same descriptions for one person often occur at different documents. If the two descriptions of a person are the same completely, only one is reserved. In addition, we also need to consider the situations that two descriptions for one person are partly identical and one is the abbreviation of the other one. For example, if one description is "俄罗斯中央银行行长(Chairman of Russia Central Bank)", whether the other one is "中央银行行长(Chairman of Central Bank)" or "行长(Chairman)". These descriptions in fact refer to the same thing, and we also only reserve one. We conduct string matching for any two descriptions. If one description is contained by the other, it will be seen as duplicate and removed.

For the same person, there are always several different names. If we can combine them into one, the description set will be more consistent and integral. This problem can be seen as a coreferential one, which doesn't involve pronouns. If one name is the fisrt name or last name of a person and there also exists the full name, we need to combine their descriptions and use the full name as the reference of the person. For example, the name "埃杰维特(Ecevit)" has the description of "土耳其副总理(Deputy Prime Minister of Turkey)" and name "比伦特·埃杰维特(Bulent Ecevit)" has "副总理(Deputy Prime Minister)" as description. We combine them and name "比伦特·埃杰维特(Bulent Ecevit)" with description "土耳其副总理(Deputy Prime Minister of Turkey)" is the final result. In addition, the results of descriptions can help to conduct person merging. For the Chinese name, often a family name plus a title noun can be used to refer to a person. For one example, the family name of "江(Jiang)" with description "主席(President)" is the same person as "江泽民(Jiang Zemin)" with description "国家主席(President of China)", while the family name of "江(Jiang)" with description "医生(doctor)" is not.

6.3.2. **Timestamping.** Descriptions will track persons' history and then time is an important factor. Thus, whenever a description is extracted, we place a timestamp on it. And every description owns the time characteristic. For example, we refer to the newswire about "Hu Jintao", who was "vice president of P.R.C" from 2002 to March 2003, elected as "president of P. R. C" from March 2003, and elected as "chairman of the Central Military Commission of P.R.C" in Sept. 2004. Here we can see that it is very useful to record all descriptions of a person at different time. With one timestamp, we can describe more precisely a person, making clear when the description is used with the person. Descriptions can be referenced and reused in appropriate places.

The extraction of time information is mainly divided into three steps.

1) Firstly, when there are no temporal phrases in the text, we directly take the date when the news is published as the time feature of a description.
2) Secondly, if one news begins with a temporal phrase, e.g. "本报/r 北京/ns [１２月/t ３０日/t]TP 讯/Ng", here the chunk tag "TP" means a temporal chunk, which is used to replace the publishing time.

For a description, if the sentence it occurs contains a temporal phrase, we suppose that the temporal phrase will modify the description, and then we take this phrase as the timestamp of the description.

7. Experimental Results. It is very difficult to evaluate the precision and recall due to the large collection of the documents involved. For small news documents sets, it is possible to manually inspect them and calculate the precision and recall. Unfortunately, as [hu 2004] said, this evaluation approach does not scale and becomes infeasible for large collection of literature. Firstly, we introduce the experimental data and distribution of manual descriptions in it and formulized definition of measures. Patterns and weighted knowledge are used to extract descriptions, and we will give some experimental results to know these resources more clearly. Finally, the results of automatically extracting descriptions are given. Our main goal is to compare the results of extracting descriptions at different stages and judge how much relevancy calculation improves the performance of extracting descriptions. In addition, the result of handling duplicates is also given to see the distribution of descriptions for every person.

7.1. **Corpus.** The annotated corpus we use is the PolyU treebank[1], half one year's news from People's Daily of the year 1998, which has been annotated by NLP tools and verified manually. There are totally 17,292 pieces of news. The annotation includes word segmentation by space, POS tags, and chunk categories. The POS tag set consists of 43 tags proposed by Peking University standard [Yu 1998], and the set of chunk category includes 21 categories, illustrated in appendix. All the person names in the corpus are tagged by the POS "/nr", through which we have found that there are about 13,020 occurrence. Because some person names occur at different places, there are in fact only 5,397 distinct person entities. In order to evaluate the system performance, we divide the entire corpus into two parts. The larger part called training corpus, contains about 5 months' newswire, which can be used to generalize the extraction patterns and improve the ontological lexicon. In the training corpus, there are about 11,339 occurrences of person names which include 4,853 distinct names.

The smaller testing corpus contains one month's news, in which 1,681 person names have occurred and there are 797 distinct people's entities. In order to evaluate the performance of description extraction, we manually identify all the descriptions in the text beforehand. 654 persons out of them have one or more than one descriptions.

[1] http://www4.comp.polyu.edu.hk/~cclab/index.shtml?p=projects_treebank&lv=2&cat=1,2,1&i=6

DDPE	Counts	DDPE	Counts
1	488	5	1
2	96	6	0
3	28	7	1
4	12	8	1

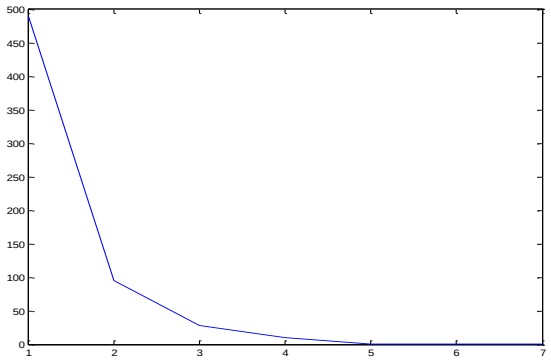

FIGURE 8. DDPE distribution in testing corpus

The numbers of descriptions before duplicate handling and after it are 1,085 and 832 respectively. After duplicate handling, there are 765 distinct persons, 627 of which have one or more descriptions and 488 have only one description. Table 3 shows the distribution of manual descriptions in the testing corpus. Although the size of testing data is not large enough, we can see that the distribution of distinct descriptions per entity (DDPE) is true of Zipf's Law from figure 8. Horizontal axis in the left figure means DDPE, and vertical axis represents the corresponding number of every DDPE.

TABLE 3. Manual Data of corpus

before duplicate handling	**Person occurrence**	**distinct persons**	**persons with description(s)**	**description occurrence**
	1,681	797	654	1,085
after duplicate handling	distinct descriptions	distinct persons	persons with description(s)	persons with one description
	832	765	627	488

To evaluate the effectiveness of our method to extract descriptions, we adopted the standard measures of precision, recall and $F_{\beta=1}$. Precision is defined to be the ratio of correct descriptions acquired divided by the total number acquired by the system. Recall is the ratio of correct descriptions acquired divided by the total number of correct descriptions in the corpus. The $F_{\beta=1}$ measure combines precision and recall with an equal weight. We can see their formal definition as in the following formula.

$$\text{Precision} = \frac{\text{the number of correct descriptions acquired}}{\text{the number of descriptions acquired}}$$
$$\text{Recall} = \frac{\text{the number of correct descriptions acquired}}{\text{the number of descriptions in the corpus}} \qquad (11)$$
$$F_{\beta=1} = \frac{2 * \text{Pr}esicion * \text{Re}call}{\text{Pr}ecision + \text{Re}call}$$

7.2. **Resources.**

7.2.1. **Patterns.** To learn rules, we collect the contexts surrounding person entities on the training corpus according to which we generalized about 10 rules finally as depicted in section 6.1. We made use of those rules to automatically extract candidate descriptions through pattern matching. We are concerned about how many correct descriptions can possibly be acquired by each rule in the rule set. *coverage* is used to measure the efficiency of a rule, defined as in formula 1. For every description in the corpus (acquired manually), if it can be covered by some rule, whether correct or partly correct, it will increase the coverage of the rule set. For example, if the context $c_1c_2\ldots\ldots c_n$ surrounding a person entity is extracted by some rule and the sequence of $c_2\ldots\ldots c_n$ is a correct description, here we still think that $c_1c_2\ldots\ldots c_n$ can increase the *coverage* of the rule. The *coverage* of a rule set gives the upper limit of recall. However, the higher the coverage of the rule set, more burden the next stage will bear. Because there are more rules, the system will collect more candidates which will be sent to the next module to pick out the correct ones.

$$coverage(R_1) = \frac{(\text{partly})\ \text{correct descriptions covered } R_1}{\text{descriptions existing in the testing corpus}} \tag{12}$$

Table 4 shows the coverage of the rules introduced in section 5.1. We can see that the number of descriptions covered by the rule set is about 1,023, and the coverage is about 0.943. As the number of correct descriptions is only 993, we can conclude that the precision is 0.90(993/1104) and the recall is 0.915 (993/1085).

TABLE 4. Coverage of rules

Rule	Candidates aquired	Correct desc.	Coverage
Rule 1	727	682	693/1085
Rule 2	332	284	302/1085
Rule 3	20	11	12/1085
Rule 4	9	4	4/1085
Rule 5	3	3	3/1085
Rule 6	1	1	1/1085
Rule 7	7	4	4/1085
Rule 8	1	1	1/1085
Rule 9	2	2	2/1085
Rule 10	2	1	1/1085
Total	1104	993	0.943

We can see that the top rules have more generalized capability. The criterion of choosing a rule is that it has a higher coverage and don't incur too many false cases. Because expressive styles of natural language are too flexible and some phenomena are very difficult to generalize into a good rule. Our rule set is not complete and can't cover all the descriptions, needing further improvement.

7.2.2. **Result of relevancy calculation.** We have downloaded the free 2000-version HowNet, which include about 1,667 sememes and 68,630 terms. Every term is explained by a set of sememes. As introduced in section 4, through the combination of SRW and semantic relevancy, we can get the contribution of sememes to being a

description for a term. Table 5 illustrates the examples of sememes with their contribution, and table 6 shows some examples of terms with their relevancy with person entities.

Because HowNet lexicon is different from the segmentation lexicon, there are a lot of terms which have no semantic definition. According to the method introduced in section 4.3, we have automatically defined for 23,058 new terms according to the initial lexicon, which include about 4,000 proper names.

TABLE 5. Sememes with contribution

Sememe	**Contr.**	**Sememe**	**Contr.**
众|mass	0.42	人|human	0.38
官|official	0.41	老|aged	0.37
男|male	0.41	家|family	0.37
体格|physique	0.39	举止|behavior	0.36
外交|diplomatic	0.38	友|friend	0.35

TABLE 6. Examples of terms with relevancy

term	**Rele**	**term**	**No.**	**term**	**No.**
人	0.35	会长	0.31	尼日利亚人	0.29
诸君	0.34	主持	0.31	首相	0.29
波斯人	0.33	长辈	0.31	秘书长	0.29
阿拉伯人	0.33	外交部长	0.31	巴基斯坦	0.28
伉俪	0.32	临时代办	0.31	伊朗	0.27
爷爷	0.31	使节	0.31	刚果	0.27
妯娌	0.31	专员	0.31	中年人	0.26
弟兄	0.31	君王	0.30	领导人	0.26
姊妹	0.31	独生子	0.30	领导干部	0.26
姐妹	0.31	官僚	0.30	钟点工	0.26

7.3. **Results of description extraction.**

7.3.1. **Results of semantic filtering.** After pattern matching, we have obtained a set of candidate descriptions. As depicted in section 6.2, relevancy of phrases are calculated to filter out those ones which have been wrongly collected. Of course it is better if relevancy calculation can help to recall the descriptions which can't be covered by crude selection. Here we conduct one experiment. That is, we compute the relevancy values of all chunks surrounding a person entity which aren't covered by patterns, and take as descriptions those chunks with higher relevancy. The result is illustrated in table 7. We can see that relevancy calculation can recall most of the descriptions which aren't recalled by patterns, however, it has also recalled more noise. Here we conclude that relevancy calculation can be better made use of with the guidance of patterns. Then our

paper mainly makes use of relevancy to implement the filtering role. That is, weighted knowledge is used to filter out wrong descriptions while assuring the correct ones aren't deleted.

TABLE 7. recalling results of relevancy calculation

	Correct descriptions recalled	**Incorrect descriptions**
Number	62	639

Because the rule set gives the upper limitation of the system's recall, we expect to improve the precision while securing the recall. For the set of candidate descriptive phrases, we conduct two experiments to verify our method of filtering.

Firstly, according to the initial lexicon of HowNet, we get the average value of SRW and semantic relevancy for every term (a simple combination) to conduct semantic filtering on the candidate descriptions. The result is given in the middle row of table 8, which shows that the performance doesn't improve much.

As depicted in section 5.3, aiming at the inadequacy of lexicon and data sparseness of corpus, we use the improved lexicon and tuned relevancy value to filter out the candidate set (called improved method). The last row in table 8 shows that this method can improve precision more while doesn't debase the recall much. The baseline in the first row in table 8 means that only crude selection is conducted.

TABLE 8. Performance results of several methods

Method	**Precision**	**Recall**	**F**
Baseline	0.897	0.915	0.906
Initial lexicon + simple combination	0.954(960/1006)	0.885(960/1085)	0.918
improved lexicon + relevancy tuning	0.972(984/1012)	0.907(984/1085)	0.938

From the comparison of F measures in table 8, we see that the combination of improved lexicon and relevancy turning only promote the performance about 3 percent. Because there is a balance between precision and recall, we put more emphasis on precision. With higher precision, the extracted descriptions can be better referenced.

7.3.2. **Result after duplicate handling.** For the descriptions extracted through our improved method, we conduct duplicates handling as introduced in section 6.3.1. We merge those names which refer to the same person, and remove those repeated descriptions and similar ones with the same reference. At last, we can get 617 person entities which have one or more than one descriptions, and the descriptions sum up to 838. For every DDPE we list the number of whose descriptions are completely correct. We can see that the precision and recall of entities with descriptions are respectively 0.908 and 0.9.

TABLE 9. Performance of duplicate handling

DDPE	counts	Correct ones	precision	recall
1	480	457	0.952	0.936
2	89	73	0.820	0.760
3	26	19	0.731	0.679
4	14	8	0.571	0.667
5	6	2	0.333	1
8	2	1	0.50	1
All	617	560	0.908	0.900

Different person entities perhaps have the same name. During duplicate handling, we consider a distinct name as a distinct entity. Then all the descriptions extracted from the same name belong to the same person. For example, a person named “Ma Xiaochun” is a reporter of Xinhua News Agency, and another person with the same name is a famous 9-Dan go player. This also brings some trouble for later reference, which will be considered in future research.

8. **Conclusions.** Descriptions for person entities are important in other applications. In this paper, we have presented a means of extracting descriptions for person entities from news corpus. A method with a pipeline of crude selection and fine filtering was applied. Extraction patterns used in crude selection are generalized through observation of annotated corpus, which provide structural information in guiding extraction. With the distribution of rules and ontological semantic knowledge, we quantified the relevancy of every term with person names in the text, which can be used to filter those inappropriate ones. Through experimental testing, results showed the method is successful, with a higher precision and a stable recall.

In future, we would like to increase the amount of patterns that the system is able to match. For the automatically improved lexicon, we will verify the definitions of those new terms and further improve the relevancy value to represent relations between terms.

REFERENCES

[1] Dragomir R. Radev. "Learning Correlations between Linguistic Indicators and Semantic Constraints: Reuse of Context-Dependent Descriptions of Entities". *Proceedings, 17th International Conference on Computational Linguistics and 36th Annual Meeting of the Association for Computational Linguistics COLING-ACL'9*8 (Montreal, Canada, August 1998).

[2] Dragomir R. Radev , Kathleen R. McKeown, Building a generation knowledge source using Internet-accessible newswire, Proceedings of the fifth conference on Applied natural language processing, p.221-228, March 31-April 03, 1997, Washington, DC.

[3] Dragomir R.Radev, Generating Natural Language Summaries from Multiple On-Line Sources: Language Reuse and Regeneration, Ph.D thesis, 1999, Columbia University, USA.

[4] Entity Detection and tracking – phrase 1, ACE Pilot study task definition, 2000, ftp://jaguar.ncsl.nist.gov/ace/phase1/edt_phase1_v2.2.pdf

[5] Schiffman, B., Mani, I., and Conception, K. Producing Biographical Summaries: Combining Linguistic Knowledge with Corpus Statistics. In Proceedings of the 39th Annual Meeting of the Association for Computation Linguistics (ACL'2001), 450- 457. New Brunswick, New Jersey: Association for Computational Linguistics.

[6] GUARINO, N., "Formal Ontology in Information Systems", *Proceedings of FOIS'98, Formal Ontology in Information Systems,* Trento, 3-15, 1998.

[7] Douglas E. Appelt and David Israel, IJCAI-99 Tutorial, Artificial Intelligence Center, SRI International.

[8] T. Nasukawa and T. Nagano, Text analysis and knowledge mining system, IBM Systems Journal: Knowledge Management, Vol.40, No.4, 2001, p967-984

[9] David D. McDonald. Internal and external evidence in the identification and semantic categorization of proper names. In Proceedings of the workshop on Acquistion of Lexical Knowledge from text, pages 32-43, Columbus, Ohio, June 1993. Special Interest Group on the lexicon of the Association for Computational Linguistics.

[10] H. Joho, Mark Sanderson, Retrieving Descriptive Phrases from large Amounts of free Text, Proceedings of the ninth international conference on Information and knowledge management,McLean, Virginia, United States, p.180 - 186, 2000

[11] Y. K. Liu.: Finding Description of Definitions of Words on the WWW. Master thesis,University of Sheffield, England, 2000. Available at : http://dis.shef.ac.uk/mark/cv/publications/dissertations/Liu2000.pdf

[12] H. Nguyen, P. Velamuru, D. Kolippakkam, H. Davulcu, H. Liu, M. Ates. Mining "Hidden Phrase" Definitions from the Web. APWeb 2003, 23-25, April 2003, Xi'an, China.

[13] H. Aholen, O. Heinonen, M. Klemettinen, and A.I. Verkanmo. : Applying Data Mining Techniques for descriptive phrase Extraction in Digital Collections, Proceedings of ADL'98, Santa Barabara, USA(4, 1998).

[14] H. Joho, Y.K. Liu, M. Sanderson, Large scale testing of a descriptive phrase finder, in the proceedings of the 1st HLT (Human Language Technologies) Conference, pages 219-221, 2001.

[15] Luke, S., L. Spector, and D. Rager. Ontology-Based Knowledge Discovery on the World-Wide Web. In Working Notes of the Workshop on Internet-Based Information Systems at the 13th National Conference on Artificial Intelligence (AAAI96). A. Franz and H. Kitano, Eds. AAAI Press, 1996, 96-102.

[16] Maedche, A.D., Ontology Learning for the Semantic Web, Norwell, Massachusetts, Kluwer Academic Publishers, 2003: http://ltc.ehb.be/terminography/presentations/pres_ceusters_smith.pdf

[17] Dong Zhendong, Dong Qiang, *HowNet,* http://www.keenage.com

[18] Roberto Navigli, Paola Velardi, Aldo Gangemi: Ontology Learning and Its Application to Automated Terminology Translation. IEEE Intelligent Systems 18(1): 22-31 (2003).

[19] Eivind Bjoraa (2003). Ontology Guided Financial Knowledge Extraction from Semi-structured Information Sources. Thesis in Master of Technology Degree Information and Communication Technology. Agder University College.

[20] Nirenburg S., Marjorie M. and Stephen B. (2003). Enhancing Recall in Information Extraction

through Ontological Semantics. In Proceedings of the Workshop on Ontologies and Information Extraction. Bucharest, Romania, August 2003.

[21] C.-W. Wu and C.-L. Liu. Ontology-based text summarization for business news articles, Proceedings of the ISCA Eighteenth International Conference on Computers and Their Applications (CATA'03), 389-392. Honolulu, Hawaii, USA, 26-28 March 2003.

[22] G. A. Miller. Wordnet: A Lexical Database. Communications of the ACM, 38:11:39--41, 1995.

[23] Yu Shiwen, et al. 1998. The Grammatical Knowledge-base of contemporary Chinese: a complete specification. Tsinghua University Press, Beijing, China.

[24] Hu X. Lin T.Y, Song I-Y. Lin X. Yoo I.. Lechner M., Song, M., Ontology-based Scalable and Portable Information Extraction System to Extract Biological Knowledge from Huge Collection of Biomedical Web Documents, the 2004 IEEE/ACM Web Intelligence Conference, Sept, 2004.

[25] Liuqun, Li Sujian, Lexical semantic similarity computation based on HowNet, Computational Linguistics and Chinese Language Processing, Vol.7, No.2, August 2002, pp.59-76.

[26] LI Sujian, ZHANG Jian, HUANG Xiong, BAI Shuo, and LIU Qun,Semantic Computation in a Chinese Question-Answering System, Journal of Computer Science&Technology(JCST),Vol.17, No.6, 2002.

[27] Douglas E. Appelt, David Israel, Introduction to Information Extraction Technology, IJCAI-99 Tutorial,August 2, 1999, Stockholm, Sweden.

APPENDIX

1 POS TAG SET

ag adjective morpheme	l habitual word	r pronoun
a adjective	m numeral	s location word
ad adverb-adjective	ng noun morpheme	t time
an adnoun	n noun	u auxiliary
b distinguished word	nr person name	vg verb morpheme
c conjunction	ns toponym	v verb
d adverb	nt organization proper	vd adverb-verb
e exclamation	noun	vn gerund
f position word	nx foreign character	w punctuation
h heading element	nz other proper noun	x unknown word
i idiom	o onomatopoeia	y modal word
j abbreviation	p preposition	z state word
k tail element	q quantifier	

2 CHUNK TAG SET

NP noun phrase	DP adverb phrase	BA BA phrase
TP time phrase	PP proposition phrase	BEI BEI phrase
FP positon phrase	QP quantifier phrase	RP pronoun phrase
VP verb phrase	DE DE phrase	
AP adjective phrase	SU SUO phrase	