Journal of Industrial Information Integration 3 (2016) 31–36

Contents lists available at ScienceDirect

Journal of Industrial Information Integration

journal homepage: www.elsevier.com/locate/jii

Semantic input method of Chinese word senses for semantic document exchange in e-business

Guangyi Xiao[a,*], Jingzhi Guo[b], Zhiguo Gong[b], Renfa Li[a]

[a] College of Computer Science and Electronic Engineering, Hunan University, Changsha, China
[b] Faculty of Science and Technology, University of Macau, Macau, China

ARTICLE INFO

Article history:
Available online 22 July 2016

Keywords:
Semantic input method
Word sense representation
Word sense disambiguation
Semantic document exchange
E-marketplace

ABSTRACT

In e-marketplace, semantic document exchange is a methodology of providing exchangeable semantic documents, which ensures document writer, writer's computer, reader's computer and document reader to share a same understanding in meaning on any exchanged document, that is, a semantic document is exchangeable across any heterogeneous contexts. This paper illustrates two kinds of semantic input method for Chinese word senses such as Word-based word sense input and Sentence-based word sense input. These two kind of word sense input methods are based on the statistic word sense representation and disambiguation. The prototype system shows both our Word-based and Sentence-based word sense Pinyin methods are promising in the text edit system. These two kinds of Chinese word sense input methods are designed for our semantic document edit: syntactic file and semantic file aliment system, designed for the semantic document exchange for e-business.

© 2016 Elsevier Inc. All rights reserved.

1. Introduction

From the dawn of Internet, electronic marketplace (e-marketplace) plays an important role in electronic business (e-business) [1,2], it is defined as the "places where buyers and sellers conduct transactions by electronic means" [3].

E-business transactions are done through a sequence of e-business functions or activities, for example, inquiring a product, making an offer, accepting an offer, making a shipping notice, and billing a payment [4]. The core of designing these functions for e-business is unambiguous document delivery. It is important to deliver an unambiguous document in meaning from a sender of a system to a receiver of another system, ensuring the identical sense-making between document sender and receiver [5].

A semantic document [6,7] is a document that is represented in the form of readable and understandable by both human and computer. An exchangeable semantic document is a document that is consistently readable and understandable not only by both computer and human of a same context but also by both document writers and document readers of different places that may have various backgrounds of meaning interpretation. By this distinction, semantic document exchange is a methodology of providing exchangeable semantic documents, which ensures document writer, writer's computer, reader's computer and document reader to share a same understanding in meaning on any exchanged document, that is, a semantic document is exchangeable across any heterogeneous contexts.

However, this document exchange between different contexts results in four important research problems [7], which are:

(1) Syntactic document structure consistency problem. A syntactic document structure is a grammatical relationship between the terms of a document on how a set of terms are organized in a grammatical pattern.
(2) Semantic document structure consistency problem. A semantic document structure is a conceptual relationship between the terms of a document on how a set of terms are organized in a semantic pattern.
(3) Semantic term consistency problem. Any terms used in a document shall be semantically consistent between contexts.
(4) Cognition consistency problem. The document writer in a context has correct semantic term meaning and correct semantic term relations in his/her mind but he/she cannot correctly write them down. For example, the document knows "orange juicer" means a juicer for squeezing orange juice, but he/she has no method to clearly write down because the document writer writes down his/her document through the input from the keyboard and mouse.

The problems illustrated in the above, are non-trivial for conducting e-business. They pose a great challenge in solving the general research problem of semantic document exchange on how to

* Corresponding author.
E-mail address: guangyi.xiao@gmail.com (G. Xiao).

http://dx.doi.org/10.1016/j.jii.2016.07.002
2452-414X/© 2016 Elsevier Inc. All rights reserved.

disambiguate the document senses made by the document writer of one context and the document senses read by document reader in another context.

A key technique is use globally identified and collaborative exchanged common vocabulary to support the semantic document editing and exchange. These common vocabularies are well design in multilingual and well disambiguated in concept level [7–9]. How to input these common vocabularies to text editor in word sense level is not trivial for semantic document exchange. This paper focus on the Chinese word senses input problem for these common vocabulary.

Pinyin-based method automatically converts Pinyin to Chinese character. Here, we use Pinyin-based method automatically convert Pinyin to Chinese word senses, not only Chinese characters but also the identifier of the word sense. But, there are only 406 syllables; they correspond to over 6000 common Chinese characters. Therefore, it is very difficult for system to select the correct corresponding Chinese word senses automatically. A higher accuracy may be achieved using a sentence-based input [10]. Sentence-based input method chooses word senses by using a language model base on context. Therefore, its accuracy is higher than word-based input method. In this paper, all the technology is based on sentence-based input method, but it can easily have adapted to word-input method.

In our approach, we use statistical language model to achieve very high accuracy. We designed an approach to Chinese language modelling for word sense representation and disambiguation. This approach enhances word senses disambiguated on tanning data to segment words, select word senses, and filter the training data.

The organization of this paper is as follows. In the second section, we briefly discuss the Chinese language model which is used by sentence-based word senses input method. In the third section, we introduce our prototype system implement of our word sense input method. In the fourth section, we propose a spelling model for English, which discriminated between Pinyin and English. Finally, we give some conclusion.

2. Word-based word sense input method

A Pinyin input is the most popular form of text input in Chinese. Basically, the user types a phonetic spelling with optional spaces, like:

maijia

And the system converts this string into Chinese word sense *WS*, consists of a tuple of a string of Chinese word *W* and a word sense identifier *id*, like:

WS = *(卖家(the seller,0 × 123))*

A full Pinyin input method chooses the probable Chinese word according to the context. In our system, statistical language model is used to provide adequate information to predict the probabilities of hypothesized Chinese word sense.

In the conversion of Pinyin to Chinese word sense, for the given Pinyin *P*, the goal is to find the most probable Chinese word sense *WS*, so as to maximize Pr(*WS* | *P*). Using Bayes law, we have:

$$\widehat{WS} = argmax_{WS}\frac{\Pr(P|WS)\Pr(WS)}{\Pr(P)} \tag{2.1}$$

The problem is divided into two parts, typing model Pr(*P* | *WS*) and language mode Pr(*WS*).

Conceptually, all *WSs* are enumerated, and the one that gives the largest Pr(*WS, P*) is selected as the best Chinese word sense. In practice, some efficient methods, such as Viterbi Beam Search [11], will be used.

The Chinese language model in Eq. (2.1), Pr(*WS*) measures the priori probability of a Chinese word sense. Usually, it is determined by a statistical word sense disambiguation model (SWSDM), such as Word Sense Representation and Disambiguation model. Pr(*P*| *WS*), called typing model, measures the probability that a Chinese word sense *WS* is typed as Pinyin *P*.

Usually, w(*WS*) word of the word sense is the combination of Chinese characters, it can decompose into $C_1, C_2, \ldots, C_n$, where c_i can be Chinese character. So the typing model can be rewritten as Eq. (2.2).

$$\Pr(P\ |WS) = \prod_{i=1}^{n}\Pr(P_{f(i)}\ |C_i\,|),\ \ W(WS) = (C_1,\ \ C_i, \ldots C_n) \tag{2.2}$$

where, $P_{f(i)}$ is the Pinyin of w_i, and W(*WS*) is a sequence of Chinese characters.

The most popular statistic word sense disambiguation model is so called Word Sense Representation and Disambiguation Model. We adapted the unified model of the word sense representation and disambiguation in next subsection.

2.1. Adapted word sense representation and disambiguation

Here, we preform knowledge-based word sense disambiguation for training data on an all-words setting, i.e., we will disambiguate all the content words in a sequence. Formally *S* is a sequence of words $(w_1, w_2, \ldots, w_n)$, and we will identify a mapping M from words to senses such that $M(i) \in \text{Senses}_{AHD}(w_i)$, where $\text{Senses}_{AHD}(w_i)$ is set of Chinese word senses encoded in the AHD for word w_i.

Therefore, given a sequence of training words $W = (w_1, w_2, \ldots, w_T)$, and we will identify a mapping M from words to senses such that $M(i) \in \text{Senses}_{AHD}(w_i)$, where $\text{Senses}_{AHD}(w_i)$ is set of Chinese word senses encoded in the AHD for word w_i. The priori probability of a Chinese word sense is computed as Eq. (2.3).

$$\Pr(WS) = \frac{1 + |W'|}{T},\ \ W' = \{w_i \in W,\ \ M(i) = WS\} \tag{2.3}$$

where *W* can also be represented as a sequence of sentences $S_1, S_2, \ldots, S_n$. and *W*' is a subset of *W* where the mapping of the word w_i after word sense disambiguation is the exactly the same with *WS*.

We extend the word sense representation and disambiguation method [12] as four-stage process:

(1) Initializing word vectors and sense vectors. Given large amounts of text data, we first use the Skip-gram model, a neural network based language model, to learn word vectors. Then we assign vector representations for sense based on their definitions.
(2) Performing word sense disambiguation. Given word vectors and sense vectors, we propose simple and efficient WSD algorithms to obtain more relevant occurrences for each sense.
(3) Learning sense vectors from relevant occurrences. Based on the relevant occurrences of ambiguous words, we modify the training objective of Skip-gram to learn word vectors and sense vectors jointly. Then we obtain the sense vectors directly from the model.
(4) Finally, performing word sense disambiguation again. Based on more accurate word sense vectors, we perform the word sense disambiguation on step 2 again. Then we can calculate the priori probability of a Chinese word sense from the mapping M(*i*) directly.

2.2. Using spelling correction

In traditional statistic Pinyin-to-characters conversion system, Pr($P_{f(i)}$ | C_i), as mentioned in Eq. 2.2, is usually set to 1 if $P_{f(i)}$ is

Download English Version:

https://daneshyari.com/en/article/4973061

Download Persian Version:

https://daneshyari.com/article/4973061

Daneshyari.com