上海交通大學

SHANGHAI JIAO TONG UNIVERSITY

学士学位论文

THESIS OF BACHELOR

论文题目：基于模糊理论的语义学模型

学生姓名：李方圻

学生学号：515141910036

专　　业：英语（金融商务实务）

指导教师：王哲希

学 院(系)：外国语学院英语系

Submitted in total fulfillment of the requirements for the degree of Bachelor in English Literature and Business

A Semantic Model based on Fuzzy Theory

Fangqi Li

Advisor

Zhexi Wang

Depart of English, School of Foreign Languages

Shanghai Jiao Tong University

Shanghai, P.R.China

June. 5th, 2019

Shanghai Jiao Tong University
学位论文原创性声明

本人郑重声明：所呈交的学位论文，是本人在导师的指导下，独立进行研究工作所取得的成果。除文中已经注明引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写过的作品成果。对本文的研究做出重要贡献的个人和集体，均已在文中以明确方式标明。本人完全意识到本声明的法律结果由本人承担。

学位论文作者签名：_____

日　期：_____年 _____月 _____日

Shanghai Jiao Tong University
学位论文版权使用授权书

本学位论文作者完全了解学校有关保留、使用学位论文的规定，同意学校保留并向国家有关部门或机构送交论文的复印件和电子版，允许论文被查阅和借阅。本人授权上海交通大学可以将本学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存和汇编本学位论文。

本学位论文属于

保　密 □，在 _____ 年解密后适用本授权书。

不保密 □。

（请在以上方框内打✓）

学位论文作者签名：_____　　指导教师签名：_____

日　期：_____年 _____月 _____日　　日　期：_____年 _____月 _____日

基于模糊理论的语义学模型

摘　要

本文提出了一种基于模糊理论的语义学模型，该模型可以基于语料库显式地对于某种语言中词汇的语义进行量化，并继而使得一些语义学中的实际问题得到回答，譬如：判定词汇间的上下义词关系、判定词汇间的同义词关系等等。

模糊理论中的许多成分反映在本文提出的语义模型的众多层面中，譬如模糊集合论被用来建模论域，而模糊逻辑论和模糊关系理论被用在模型中的组合和推导部分。作为传统逻辑的一种泛化，模糊理论通过显式地处理模糊性和不确定性给自然语言处理带来了更大的张力，从而使得模型更贴近现实。应用模糊理论也使得本文提出的模型和传统语义模型相比，具有一定程度上更高的鲁棒性和可实践性。

本文提出的模型还融会贯通了现代计算语言学、特别是语义学中两个看似区别很大的分支：基于句式的上下义词挖掘和词向量提取。前者旨在利用句式资料自动化地从语料库中挖掘上下义词关系，后者旨在对词汇进行编码使得语义上的距离被反映为几何上的距离。本文中提出的模型是这两种流行的自然语言处理之综合和泛化，即可以通过调节本文模型中的参数或者采取某些近似手段得到上述两种模型。这种模型间的关联意味着本文设计的模型是一种更宽泛、更抽象的语义表达方式。

关键词： 语义学　模糊理论　自然语言处理

A SEMANTIC MODEL BASED ON FUZZY THEORY

ABSTRACT

In this paper, we propose a fuzzy theory-based semantic model, which can quantitively infer the semantic information from a corpus and solve some practical problems. Some examples are detecting hypernym/hyponym pairs or detecting synonyms.

The fuzzy theory is reflected in our model from multiple perspectives, e.g. fuzzy set theory is adopted to model the universe of discourse while fuzzy logic relationship theory is invoked to complete the inference within. As a generalization of ordinary logic, fuzzy theory grants linguistic models the ability to handle vagueness and uncertainty in natural language explicitly. Therefore the proposed model is relatively more robust compared with traditional ones.

The proposed model also unifies two distant branches in modern natural language processing techniques, namely pattern-based hypernym/hyponym detection and word embedding. The pattern-based hypernym/hyponym detection can automatically learn hypernym/hyponym pairs from a corpus, while the word embedding aims at encoding words so the semantic distance can be reflected by geometrical distance. The proposed model is a shared generalization of these two models, i.e.they can be readily obtained from our proposal by modifying parameters or adopting approximations. This fact indicates that we have proposed a more abstract and general semantic model.

KEY WORDS: Semantics, fuzzy theory, natural language process (NLP)

Contents

Chapter 1 Introduction: Current Trends in linguistic studies

Denotation and **connotation** are two equivalent ways of identifying a concept in Aristotle logic, or classical logic. In classical logic, every concept is assigned a collection of entities from the observable world. The denotation of a concept is the assigned collection of entities, while the connotation of a concept is the collection of attributes shared by its denotation. For example, let the concept be *red*, then its denotation consists of ripe apple, blood, fire truck, etc, and its connotation is the color red. Any concept can be addressed identically by its denotation and connotation, however, this duality reflects different ways of understanding the world.

Having witnessed the development and popularity of the philosophy of science, modern science and engineering are picking up the denotative view more readily than they were centuries ago (Irving, 1978). It is the confidence that human rationality is capable of exploring and inducing attributes of any concept that underlies the connotative reasoning. However, this confidence has been suffering from restless attempts that aim at identifying the boundary of rationality. Therefore we should make ourselves feel comfortable about the fact that mankind can hardly unfold the connotation of a concept comprehensively. Instead, we should resort to the denotative way of reasoning. Moreover, a concept defined from the observable reality almost always enjoys a finite or countable set as its denotation, but such properties need not be valid for its connotation.

Linguistics, bloomed within the last centuries, also undertakes this alternation in views. The attention of linguists is partially transferring from the **intrinstic grammar rules** to **natural corpus**. To draw an analogy, the study of languages is more of physics than mathematics. Mathematicians build up their formal systems out of nothing, while physicists and linguists have to modify their theories to explain the established world, which might be less elegant and concise than the coined theories.

The most renowned theories in classical linguistics such as generative syntax, universal grammar, and modal logic take the ability of formalization and logic itself, or the **langue** as granted. It turns out that their derivatives such as paraphrase algorithm (Irving, 1978) and syntax tree are somehow fragile in natural scenarios, despite their elegance and self-consistency. The reason behind this incompatibility is the fact that natural language includes numerous vague and fuzzy elements. The construction of a natural language is not deductive (compared with formal languages as mathematics, Turing machine or other programming languages), but is rather stochastic and randomized. Therefore formal logic and set theory, which was developed to underscore formal languages are insufficient to explain all natural language phenomenon.

This line of reasoning has deviated the linguistic studies from generative rules to the natural corpus. By modestly absorbing neuroscience, sociology, psychology, statistics and partially abandoning the analytic formal logic, modern linguistics successfully addressed assorted underlying patterns and information in natural scenarios. Spearheaded by Natural Language Processing (NLP), the statistics-based linguistics is a crucial element in the current trend of artificial intelligence. Moreover, NLP is one of the most competitive candidates in achieving **strong artificial intelligence**. But statistics-based linguistics suffers from the balance

between mathematical statistics and linguistic professional knowledge. Intuitively, applying statistics model to natural corpus without consulting linguists cannot guarantee that the derived conclusions are linguistically plausible and valid. NLP is now overwhelmed with ad hoc, arbitrary and empirical methods, those methods, although significantly outperform old-fashioned deeds, lack a uniform theory framework. To summarize, benchmark modern linguistic studies should combine the fundamental theories from logic and old-fashioned linguistics, together with advanced computing methods. And they should be able to unify seemingly different language models and methods into one.

In this paper, we use the methods of fuzzy theory to deal with a semantic task: identifying hypernym and hyponym. This problem is of vital importance as it reflects machine intelligence in the level of understanding the relationship between words without a supervisor. One representative application of machine intelligence is how well an algorithm can simulate human being' s ability to understanding by answering the question as *is an apple one kind of fruit?*, *is Shanghai a city?* or *what is an apple?*, *what is Shanghai?* Whose answers are nothing more than straightforward derivations of hypernym/hyponym relationships. Apart from intelligence test, these relationships can help to promote the efficacy of services based on natural language. Since hypernym/hyponym can help to **ontologize** words (i.e. to understand a word as its generalization), it is possible to provide more general and a broader range of feedback when it is applied in search engines and other interfaces.

The fuzzy theory was proposed to provide a better description of human reasoning than traditional logic (Zimmermann, 2011). It is capable of modeling the vagueness and uncertainty within natural language (Ma, 2011)(Khoury, 2007). For which reason it has found wide application in theory as well as industry. The reason why we introduce fuzzy theory to handle the problem of hypernym/hyponym is that hypernym/hyponym was initially understood as an ordinary logic relationship, but ordinary logic meets hinder in natural settings. Thus we resort to fuzzy logic tools to obtain more robust and reasonable results.

Since there have been various ad hoc proposals concerning natural language processing around hypernym detection, another aim of this study is to find a proper interpolation between them, especially the state-of-art pattern-based hypernym detection and the word embedding, a famed technique in semantics. Specifically, our proposal should be able to degenerate gracefully to these two seemingly distinct methods.

The contributions of this paper are three-folded:

1. A fuzzy theory-based semantic model is proposed to deal with the innate vagueness and uncertainty in natural languages.
2. The proposed model can be readily applied to an interactive large-scale corpus, especially search engine, hence scale well to big data.
3. The ordinary pattern-based hypernym/hyponym detection algorithm and word embedding technique can be derived as a specialized form of our model. Thus our proposal addresses the unity between these two algorithms.

The rest of this paper proceeds as follows: Chapter 2 reviews the problem of hypernym vs. hyponym and fuzzy theory, Chapter 3 is devoted to the detailed formulation of this problem and our proposal, in Chapter 4 our proposal is discussed with more practical applications and some important remark is drawn, Chapter 5

concludes the paper.

The symbols we use throughout this paper are summarized in 1–1:

Table 1–1 Summary of formal notations.

Notation	Meaning
$\mathcal{U}$	The universe of discourse, with element $e \in \mathcal{U}$.
$\mathcal{X}_{\mathcal{F}}(\cdot)$	$\mathcal{X}_{\mathcal{F}}(\cdot) : \mathcal{U} \to [0, 1]$ is the indicator function of (fuzzy) set $\mathcal{F}$.
$\mathcal{V}$	The vocabulary, i.e., the collection of words.
$u \to v$	The proposition that u is the hypernym of v, with $u, v \in \mathcal{V}$.
$\mathcal{H}$	The hypernym/hyponym relationship, $\mathcal{H} \subset \mathcal{V} \times \mathcal{V}$.
$\mathcal{D}$	A corpus $D = \{S_1, S_2, \cdots\}$, where S_i is a sequence of words.
$[\cdot]$	A placeholder.
$S[\cdot]$	A context, it is a finite sequence of words with one and only one placeholder.
$p[\cdot, \cdot]$	A pattern, a finite sequence of words with two and only two placeholders.

Chapter 2 Review of Background Knowledge

2.1 Hypernym vs. Hyponym

For a given language, we denote its vocabulary, i.e., the collection of its words by $\mathcal{V}$. The hypernym-hyponym relationship is a binary semantic relationship with words, denoted it by $\mathcal{H} \subset \mathcal{V} \times \mathcal{V}$. According to the definition from Wikipedia, a word u is the hypernym of another word v (v is the hyponym of u) if the **semantic field** of u covers that of v, such a pair is denoted by:

$$u \to v. \tag{2–1}$$

Using the dictions from Chapter 1, u is the hypernym of v if the denotation of v is a subset of the denotation of u, i.e., the connotation of u is a subset of that of v. For example, let u be *books* and v be *dictionaries*, then u is the hypernym of v (v is the hyponym of u) since every dictionary is necessarily a book, i.e.,

$$books \to dictionaries.$$

Since the denotation of *dictionaries* is included in that of *books*. As a specific category of book, dictionaries enjoy other properties such as constraints on their contents, thus the connotation of *books* is a subset of that of *dictionaries*.

However, what we are interested in is to detect the hypernym/hyponym relationship from a natural corpus, whose language might be unfamiliar to us. From the statistical perspective, a language should be taken as a self-consistent symbolic system to be explored, and it does not necessarily have real-world referents. Therefore it is not always appropriate to query into the denotations of concepts and infer their inclusion order. Given a corpus $\mathcal{D}$, a collection of sentences, where each sentence is a finite sequence of words, we are to extract or propose an algorithm that can automatically extract the corresponding set of hypernym/hyponym pairs $\mathcal{H}$. This task has been done for English by WordNet, which provides numerous hypernym/hyponym pairs. WordNet is maintained manually and is therefore robust but expensive.

So far, the extraction of hypernym/hyponym from a corpus has been pattern-oriented (Snow, 2005). Considering the sentence *...apple, and other fruits...*, it is clear that *fruit* is the hypernym of *apple*. So it can be inferred that whenever there exists a sentence *...v, and other u...* in $\mathcal{D}$, we can readily conclude $u \to v$ and put (u, v) into $\mathcal{H}$. Most pattern-oriented extraction methods of hypernym/hyponym pairs from corpus consist of three steps:

1. Propose sentence patterns that grasp hypernym/hyponym pairs. This is usually done by professionals.
2. Extract hypernym/hyponym pairs from the corpus. This can be readily done by regular expression (RegEx) matching.
3. Eliminate illegal pairs. This is again usually done by professionals or predefined rules.

This procedure can be formulated as Algorithm 2–1:

Algorithm 2–1 Learning hypernym/hyponym relationships from a corpus with patterns.

Input: The vocabulary of the language $\mathcal{V}$;
The collection of hypernym/hyponym patterns $\mathcal{P} = \{p[\cdot, \cdot]\}$,
each $p[\cdot, \cdot] \in \mathcal{P}$ is a sentence with two placeholders;
The corpus $\mathcal{D}$, each $s \in \mathcal{D}$ is a finite sequence of words;
The compared words u and v, $u, v \in \mathcal{V}$;

Output: Whether u and v form a hypernym/hyponym pair.

flag f = False;
for p in $\mathcal{P}$ **do**
  for s in $\mathcal{D}$ **do**;
    if $s = p(u, v)$ **then** f=True;
  end for
end for
return f.

Where $p(u, v)$ is a sentence constructing by plugging words u and v into the placeholders of p. For example, let p be

[], *one kind of* [],

then $p(u, v)$ reads

u, one kind of v.

There have been fruitful studies on this topic, including the extraction of hypernym/hyponym pairs for languages other than English (Sumida, 2008)(Sang, 2007), the proposal of more sophisticated filter rules (Caraballo, 1999)(Ritter, 2009), and the automatic extraction of sentence patterns (Snow, 2005). But most of the proposed methods suffer from the arbitrariness of professionals and the limitation of individual language, moreover, they lack insightful observation into the logic essence of the hypernym/hyponym relationship. There are also methods that do not dependent on the pattern (Navigli, 2010)(Yamada, 2009), but they failed to become well recognized.

After selecting a trial set of hypernym/hyponym relationships, some algebraic tricks are utilized to explore more.

In order to simplify the further discussion, assume that we are equipped with two granted patterns, namely A-type and O-type categorical propositions:

A-type: *all* [·] *is* [·],
O-type: *some* [·] *is not* [·].

They two can readily help to form a bound of hypernym/hyponym (with respect to the order defined by inclusion) from a corpus $\mathcal{D}$. To begin, we formulate hypernym/hypohym as an ordinary subset, i.e., $\mathcal{H}$

is a subset of $\mathcal{V}^2$. Any pair $(u, v) \in \mathcal{H}$, which reads $u \to v$ is directly supported by an A-type categorical proposition in $\mathcal{D}$ readed *all v are u*. Therefore by collecting all A-type categorical propositions from $\mathcal{D}$ (pattern-based filters can readily turn a sentence with a characteristic pattern into an A-type categorical proposition and vice versa), we obtain a trial collection of hypernym/hypohym pairs:

$$\mathcal{H}_1 = \{(u, v) : u, v \in \mathcal{V}, \textit{all v are u} \in \mathcal{D}\}. \tag{2–2}$$

This trial collection can be readily expanded using the fact that $u \to v$ and $v \to w$ imply $u \to v$, which is straightforward from the definition of hypernym/hypohym. Following this line of reasoning, a word u is the hypernym of v if there exists a finite sequence of words $\{w_i\}_{i=1}^N$ such that $u \to w_1, w_N \to v$ and $\forall i = 1, \cdots, N-1, w_i \to w_{i+1}$. This procedure is essentially a path locating on a directed graph introduced by $\mathcal{H}_1$, i.e., $(u, v) \in \mathcal{H}_1$ denotes an edge from the vertex u to another vertex v. Completing the paths results in an expanded collection of hypernym/hypohym pairs:

$$\mathcal{H}_2 = \{(u, v) : \text{there is a path from } u \text{ to } v \text{ in the graph introduced by } \mathcal{H}_1\}. \tag{2–3}$$

It is necessarily true that $\mathcal{H}_1 \subset \mathcal{H}_2$. Note as well that $\mathcal{H}$ is upper bounded (with respect to the order introduced by inclusion) by $\mathcal{H}_3$ with:

$$\mathcal{H}_3 = \mathcal{V}^2 - \{(u, v) : \textit{some v is not u} \in \mathcal{D}\}. \tag{2–4}$$

Therefore by identifying all categorical propositions of A-type and O-type, the hypernym/hypohym relationship can be bounded between $\mathcal{H}_2$ and $\mathcal{H}_3$, formally:

$$\mathcal{H}_2 \subset \mathcal{H} \subset \mathcal{H}_3. \tag{2–5}$$

The bound can be embarrassingly untight due to the limited size of $\mathcal{D}$ or the failure of detecting A-type and O-type categorical propositions. As an extreme example, consider the case where $\mathcal{D} = \emptyset$, then:

$$\mathcal{H}_1 = \emptyset, \mathcal{H}_2 = \emptyset, \mathcal{H}_3 = \mathcal{V}^2.$$

And the bound by (2–5) reduces to the trivial one $\mathcal{H} \subset \mathcal{V}^2$. However, be the corpus large enough, there remains problems with the formulation, i.e., pattern plus algebraic expansion, developed so far:

1. Detecting of characteristic pattern or categorical proposition requires professional knowledge, which might be unavailable in many scenarios (e.g., pattern recognition from ancient languages.)
2. The hypernym/hypohym relationship is reflected in ways far richer than mere categorical propositions. In fact, one is capable of inducing $u \to v$ without querying about the accuracy of *all v are u*. Ignoring these effects might lead to a poor boundary for $\mathcal{H}$. Hitherto almost all hypernym/hypohym relationship detection algorithms ignore its reflections between sentences.
3. The natural corpus contains imprecision and mistakes, and some indeliberate errors might be fatal for the formulated method. For example, a indeliberate contradiction *all v iu v* with *some v is not u* might make (2–5) an illegal relationship and leave $\mathcal{H}$ intractable.

In light of the listed problems, we propose a semantic model that utilizes fuzzy theory, which is independent of specific language, capable of absorbing inter-sentence information and robust against errors.

2.2 Fuzzy Theory

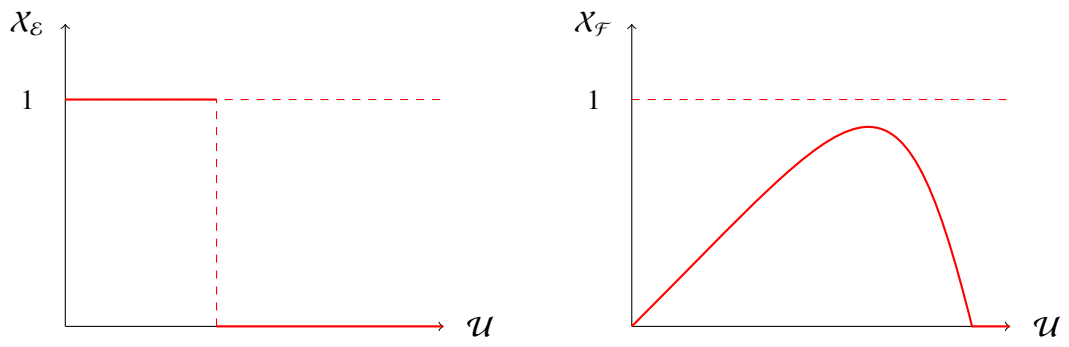
The fuzzy theory consists mainly of fuzzy set theory, fuzzy logic, and fuzzy cybernetics, etc (Zimmermann, 2011). It is a deviation from the ordinary set theory and the ordinary logic. For a subset $\mathcal{E}$ of the universe of discourse $\mathcal{U}$ ($\mathcal{U}$ represents the universal set of referents), an element e of $\mathcal{U}$ either belongs to $\mathcal{E}$ or not according to classical set theory, which is reflected by the indicator function of $\mathcal{E}$:

$$\mathcal{X}_{\mathcal{E}}(e) \in \{0, 1\}, \text{ where } e \in \mathcal{U}. \tag{2–6}$$

The indicator function $\mathcal{X}_{\mathcal{E}}(\cdot) \in \{0, 1\}^{\mathcal{U}}$ identify $\mathcal{E}$ from the denotative perspective. Fuzzy theory, on the other hand, address a fuzzy subset $\mathcal{F}$ of $\mathcal{U}$ by the fuzzy indicator function $\mathcal{X}_{\mathcal{F}}(\cdot)$:

$$\mathcal{X}_{\mathcal{F}}(e) \in [0, 1], \text{ where } e \in \mathcal{U}. \tag{2–7}$$

An example of indicators of an ordinary set and a fuzzy set is given in 2–1:



(a) Ordinary set. (b) Fuzzy set.

Figure 2–1 The indicator functions for ordinary set and fuzzy set.

The indicator functions readily yield logic propositions. For an ordinary subset $\mathcal{E}$ featured by $\mathcal{X}_{\mathcal{E}}(\cdot)$, the statement $e \in \mathcal{E}$ is either true or false according to the value of $\mathcal{X}_{\mathcal{E}}(e)$. However, for a fuzzy subset, $\mathcal{F}$ the statement $e \in \mathcal{F}$ has a fuzzy value between zero (absolutely false) and one (absolutely true). This enables the model of noise, uncertainty and the level of confidence. Therefore fuzzy theory has found effective applications in control systems, stochastic modeling and language interpretation (Zimmermann, 2011). Particularly, there have been works that using fuzzy theory to model semantics (Bordogna, 1993)(Herrera, 2000)(Pedrycz, 2006), where nouns are interpreted as fuzzy subset indicators (Zadeh, 1977) and adjectives are interpreted as operators in the space of indicator functionals (Zadeh, 1972).

The generalization of ordinary logic to its fuzzy counterpart is one kind of interpolation. One only has to make sure that the generalization can gracefully reduce to the ordinary case in limit case. The concrete form of generalization can take various forms. We introduce several generalizations that are to be adopted in the rest of this paper. Since one and zero represent definite true and false, the confidence of a conjunction *a and b* is given by:

$$CF(a \wedge b) = \min\{CF(a), CF(b)\} = CF(a) \cdot CF(b), \tag{2–8}$$

when a is true and b is false, let $a = 1$ and $b = 0$, we have $\min\{a, b\} = 0$ hence $a \wedge b$ is false. So this generalization is plausible. And

$$CF(a \vee b) = \max\{CF(a), CF(b)\}, CF(\text{not } a) = 1 - CF(a) \tag{2–9}$$

When there holds an implication $a \Rightarrow b$, we naturally have:

$$CF(b) \geq CF(a) \tag{2–10}$$

That is to say, the confidence of b is lower bounded by that of a, this analog is in accordance with ordinary logic. At last, the confidence of a proposition is given by:

$$CF(a \Rightarrow b) = \max\{(1 - CF(a)), CF(b)\} \tag{2–11}$$

It is trivial to check that (2–9) and (2–10) are in accordance with their ordinary version.

Chapter 3 Proposed Method

In this chapter, we apply the fuzzy theory and the philosophy behind to the problem of learning hypernym/hyponym pairs from a corpus. Firstly, we generalize $\mathcal{H}$ from an ordinary subset of $\mathcal{V}^2$ to a fuzzy subset and propose corresponding methods. Secondly, we discuss practical techniques that bring the computations in the previous proposal back to the ground.

3.1 The Fuzzy Version of Hypernym/Hypohym Relationship

Returning to the ideal denotative definition of two concepts u and v, as long as u and v are innately subsets of the universe of discourse, $\mathcal{U}$, the propostion $u \to v$ is reflected by their indicator functions:

$$(\forall e \in \mathcal{U}, \mathcal{X}_v(e) = 1 \Rightarrow \mathcal{X}_u(e) = 1) \Rightarrow (u \to v), \tag{3–1}$$

which is obvious from the defition of hypernym/hypohym. (3–1) indicates that whenever something is v, it must be u as well. Thus the denotation of v is covered by that of u.

Since it has already be indicated that concepts are more properly idenfified as fuzzy subsets of $\mathcal{U}$ (Zadeh, 1972), the condition in (3–1) can be readily generalized to its fuzzy version, where we applied (2–10) to the antecedent of (3–1) and recall that the indicator function $\mathcal{X}$ is nothing but a measure of confidence w.r.t. a concept:

$$(\forall e \in \mathcal{U}, \mathcal{X}_v(e) \leq \mathcal{X}_u(e)) \Rightarrow (u \to v). \tag{3–2}$$

As far as (3–2) is a proposition in itself, it further implies (using (2–10)):

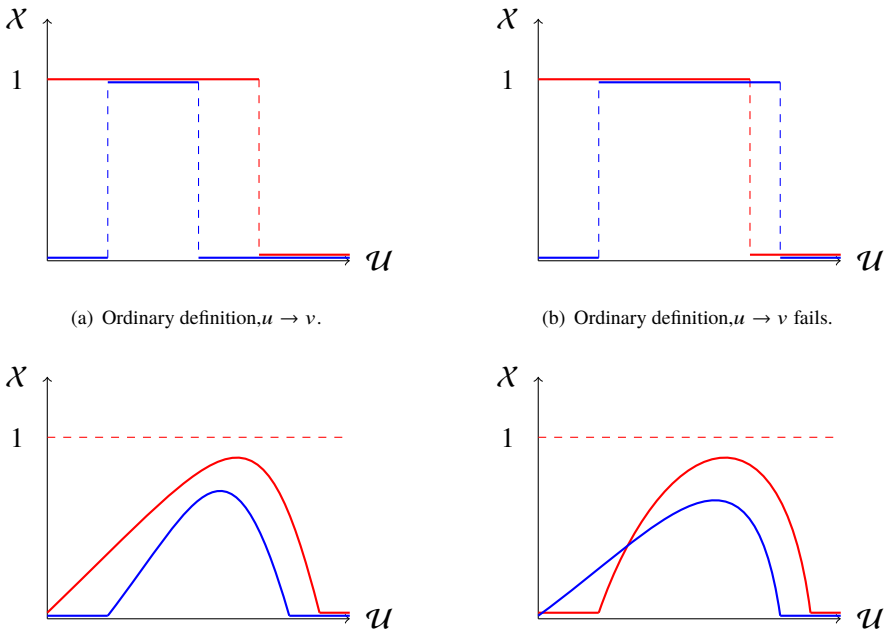
$$CF(u \to v) \geq CF(\forall e \in \mathcal{U}, \mathcal{X}_v(e) \leq \mathcal{X}_u(e)). \tag{3–3}$$

Where $CF(p)$ denotes the confidence of some proposition p. (3–3) is obtained by applying (2–10) to (3–2). We illustrate the ordinary and fuzzy version of two concepts u and v in Figure 3–1.

Figure 3–1 (a) and (b) present an ordinary A-type and O-type categorical propositions respectively. Figure 3–1 (c) is a fuzzy A-type categorical proposition, since if something e is v with confidence $\mathcal{X}_v(e)$, it is always true that e is u with a higher confidence $\mathcal{X}_u(e) > \mathcal{X}_v(e)$. Figure 3–1 (d) is a fuzzy O-type categorical proposition and is worth noting. As emphasized in Figure 3–2, the counter-example in Figure 3–1 (d) itself might not be representative referents of v, therefore the confidence of this refutation is actually very fragile. In cases as Figure 3–1 (d) or Figure 3–2, the proposition $u \to v$ should still obtain a relatively high confidence instead of zero.

To absorb the philosophy behind Figure 3–1 and Figure 3–2, the confidence function on r.h.s of (3–3) should take two indicator functions as the parameter and be of the form $[0,1]^{\mathcal{U}} \times [0,1]^{\mathcal{U}} \to [0,1]$ such that:

- If $\forall e \in \mathcal{U}, \mathcal{X}_v(e) \leq \mathcal{X}_u(e)$ then $CF(\mathcal{X}_v, \mathcal{X}_u) = 1$.



(a) Ordinary definition,$u \to v$.

(b) Ordinary definition,$u \to v$ fails.

(c) Fuzzy definition,$u \to v$ with confidence 1.

(d) Fuzzy definition,$u \to v$ with confidence less than 1.

Figure 3–1 Illustration of $u \to v$ with different confidence in ordinary and fuzzy settings. The red curve and blue curve denote $\mathcal{X}_u$ and $\mathcal{X}_v$ respectively.

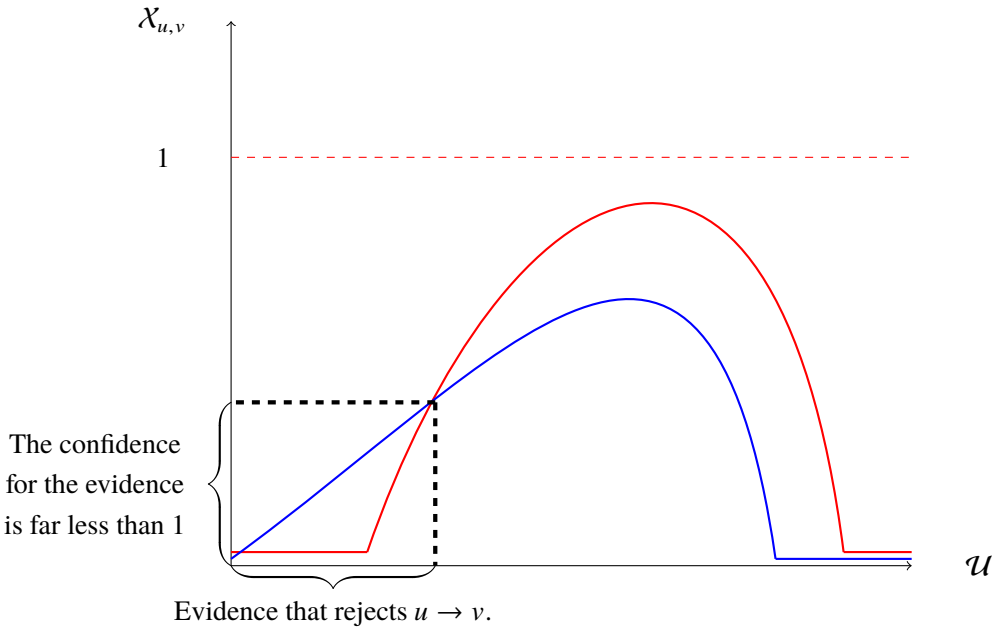


Figure 3–2 $u \to v$ with confidence less than 1.

- If $\mathcal{X}_v(e) \leq \mathcal{X}_u(e)$ failes to hold on a subset E of $\mathcal{U}$, but E is relatively small or

$$\max_{e\in E}\{\mathcal{X}_v(e)\}$$

is relatively small, then $CF(\mathcal{X}_v, \mathcal{X}_u)$ should have a value close to one.

Remark: The term max here is a mere formal gadget, if the geometry of $\mathcal{U}$ has been explored further, sup might be used instead.

In light of this setting we propose two candidates of the confidence function:

$$CF_g(\mathcal{X}_u, \mathcal{X}_v) = 1 - \frac{\int_{\mathcal{X}_v > \mathcal{X}_u} (\mathcal{X}_v - \mathcal{X}_u)\,\mathrm{d}\mu}{\int \mathcal{X}_v \mathrm{d}\mu}, \tag{3–4}$$

$$CF_l(\mathcal{X}_u, \mathcal{X}_v) = 1 - \max\left\{0, \max_{e\in\mathcal{U}}\{\mathcal{X}_v(e) - \mathcal{X}_u(e)\}\right\}. \tag{3–5}$$

Where the subscripts in (3–4) and (3–5) stand for global and local.

It is easy to check if $\forall e \in \mathcal{U}, \mathcal{X}_v(e) \leq \mathcal{X}_u(e)$ then $CF_g(\mathcal{X}_v, \mathcal{X}_u) = CF_l(\mathcal{X}_v, \mathcal{X}_u) = 1$. To check for the second requirement, consider the caes when E (where $\mathcal{X}_v(e) \leq \mathcal{X}_u(e)$ failes to hold) is relatively small (*small* can be measured only by some predefined measure μ on $\mathcal{U}$) or $\max_{e\in E}\{\mathcal{X}_v(e)\}$ is small, the integral $\int_{\mathcal{X}_v > \mathcal{X}_u} (\mathcal{X}_v - \mathcal{X}_u)\,\mathrm{d}\mu$ in (3–4) will be small as well, since

$$\int_{\mathcal{X}_v > \mathcal{X}_u} (\mathcal{X}_v - \mathcal{X}_u)\,\mathrm{d}\mu \leq \mu(E) \cdot \max_{e\in E}\{\mathcal{X}_v(e)\}.$$

Meanwhile if $\max_{e\in E}\{\mathcal{X}_v(e)\}$ is small, then $\max_{e\in\mathcal{U}}\{\mathcal{X}_v(e) - \mathcal{X}_u(e)\}$ will take a trivial positive value. The formulation (3–5) avoids assuming a measure on $\mathcal{U}$, hence considers only local information (it considers only the worst counter-example against $u \to v$). (3–5) can be derived from the form $CF\ (\forall e \in \mathcal{U}, \mathcal{X}_v(e) \leq \mathcal{X}_u(e))$ and (2–8):

$$\begin{aligned} CF\ (\forall e \in \mathcal{U}, \mathcal{X}_v(e) \leq \mathcal{X}_u(e)) &= \wedge_{e\in\mathcal{U}} \mathcal{X}_v e \leq \mathcal{X}_u(e) \\ &= \min_{e\in\mathcal{U}}\{CF(\mathcal{X}_v(e) \leq \mathcal{X}_u(e))\} \\ &= \min_{e\in\mathcal{U}}\{1 - CF(\mathcal{X}_v(e) > \mathcal{X}_u(e))\} \\ &= 1 - \max_{e\in\mathcal{U}}\{\mathcal{X}_v(e) - \mathcal{X}_u(e)\} \\ &= 1 - \max\left\{0, \max_{e\in\mathcal{U}}\{\mathcal{X}_v(e) - \mathcal{X}_u(e)\}\right\}. \end{aligned} \tag{3–6}$$

But (3–4) is still indispensible since it grasps global information rather than local one. Now given the indicator functions of two concepts u and v, the confidence of the proposition $u \to v$ can be readily computed by combing (3–3) (3–4) and (3–5):

$$CF(u \to v) = CF_s(\mathcal{X}_u, \mathcal{X}_v), \tag{3–7}$$

where s stands for g or l, and we replace the inequality with equality if no refutation arises. (3–7) suggests that the only left difficulty is to infer $\mathcal{X}_u$ of a word u from a corpus.

Remark: The proposal of the global metric (3–4) seems to be a deviation of the orthodoxy fuzzy logic (3–3). However it is vital since orthodoxy fuzzy logic, a straightforward generalization of ordinary logic,

cares not about the geometry of the underlying set. But such point set interpretation might fail in our setting since the universe of discourse might not be equipped with the equivalent structure. The refutation of a proposition, according to the ordinary (fuzzy) logic, might be invalid if the refutation fails to hold almost everywhere, i.e.the refutation evidence might be a subset of the universe of discourse with measure zero. To put it into other words, we know too few about the measure on the universe of discourse, thus we should better explicitly introduce such a measure in the form as (3–4).

For example, let $\mathcal{U}$ be $\mathbb{R}$, the collection of all real numbers and let $\mathcal{X}_u(\cdot)$ be the indicator function of $[0, 3]$, $\mathcal{X}_v(\cdot)$ be the indicator function of $[1, 2] \cup \mathbb{N}$, then obviously

$$\forall e \in \mathcal{U}, \mathcal{X}_u(e) \geq \mathcal{X}_v(e)$$

failes to hold (consider $e = 4, 5, 6, \cdots$.) But one also has to note that the set on which $\mathcal{X}_n \geq \mathcal{X}_v$ fails to hold counts for measure zero (with respect to Lebesgue measure), so we have:

$$\mathcal{X}_u \geq \mathcal{X}_v \text{ holds almost everywhere.}$$

Therefore we can hardly discuss the confidence of such a predicate proposition without introducing a measure into our formulation.

3.2 Model of the Universe of Discourse

Despite the plausible nature of (3–7), its application is intractable unless there exists an established indicator function for any given word u or v and a measure on $\mathcal{U}$. Thus if we are equipped with a method of approximating the indicator function of any word from a corpus $\mathcal{D}$ then we can combine it with (3–7) to obtain a complete framework of corpus-based hypernym/hyponym detection.

Our start point is: what effects does $u \to v$ bring to the corpus? Assume that we are equipped with a corpus so large that every subset of the denotation of any word is discriminated by some sentence in it. This corpus is denoted by $\mathcal{D}_{\text{Babel}}$ to emphasize its all-inclusive potent. If $u \to v$ and there is a sentence S that contains v in the corpus, then intuitively:

$$S[v] \in \mathcal{D}_{\text{Babel}} \Rightarrow S[\textit{some } u] \in \mathcal{D}_{\text{Babel}},$$

where $S[\cdot]$ is the sentence S with v replaced by a placeholder that needs to be plugged. We define such a pseudo-sentence with one placeholder as a **context**. To name an example, consider the hypernym proposition

$$\textit{fruit} \to \textit{apple} \text{ , and } \textit{apple is red},$$

it is necessarily true that

$$\textit{some fruit is red}.$$

This line of reasoning can be formalized and enhanced as the following theorem:

Theorem 1.
$u \to v$ *iff* $(S[v] \in \mathcal{D}_{Babel} \Rightarrow S[some\ u] \in \mathcal{D}_{Babel})$.

Proof. To prove that

$$u \to v \Rightarrow (S[v] \in \mathcal{D}_{\text{Babel}} \Rightarrow S[\text{some u}] \in \mathcal{D}_{\text{Babel}}),$$

consider any sentence S in $\mathcal{D}_{\text{Babel}}$ that contains v. S utilizes the concept v by either denotation or connotation, in either way S is still plausible after replacing v by *some* u, hence $S[\textit{some } u] \in \mathcal{D}_{\text{Babel}}$.

Conversely, if $u \to v$ fails to hold, then there exists a subset of the denotation of v that is not included in the denotation of u. By assumption there exist some sentence S about this subset, and $S[\textit{some } u]$ is semantically incorrect, hence $S[\textit{some } u] \notin \mathcal{D}_{\text{Babel}}$. By reductio ad absurdum the proof is completed. □

Utilizing the fuzzy theory to reformulate the proposition in Theorem 1, we have:

$$CF(u \to v) \geq \frac{|\{S[\textit{some } u] : S[v] \in S_v\} \cap D_{\text{Babel}}|}{|S_v|}, \tag{3–8}$$

where S_v is the collection of sentences in $\mathcal{D}_{\text{Babel}}$ that contains v, i.e., $\{S : v \in S, S \in\} \mathcal{D}_{\text{Babel}}$. The fuzzification of the form $a \Rightarrow b$ (when a and b are predicates) has various versions, we adopt the method:

$$CF(a \Rightarrow b) = \frac{\int \min\{CF(a(e)), CF(b(e))\} \, \mathrm{d}\mu}{\int CF(a(e)) \mathrm{d}\mu}.$$

As an average of soft support of a to b. Whose degeneration to the ordinary sense is obvious. The reason why we adopt this form rather than (2–11) is the same as the last remark. When $u \to v$ holds in ordinary sense, the r.h.s. of (3–8) reduce to one according to Theorem 1. When some v is not u, the confidence of $u \to v$ is measuerd by the percentage of usages of v that are not compatible with u.

It is remarkable that (3–8) is similar to (3–4) if the fuzzification occurs only in transforming (3–2) and Theorem 1 but not the definitions of u and v. That is to say, if we only apply the fuzzy theory in transforming the logic reasoning but not the definition of words.

If the measure on $\mathcal{U}$ is replaced with the counting measure on the space of contexts and the domain of indicator functions is $\{0, 1\}$, then (3–4) reduces to (3–8). This remark motivates two further issues:

1. Replacing the measure on $\mathcal{U}$ by the counting measure on the space of contexts can fulfil the general integral in $\mathcal{U}$. And $e \in \mathcal{U}$ can be readily mapped as one context, or one collection of contexts.
2. There should be a fuzzy version of $\mathcal{D}_{\text{Babel}}$, where any finite sequence of words is assigned a confidence. Sequence which is a legal sentence enjoys confidence one, while sequence that are grammatically or semantically wrong has less confidence, but not necessarily zero.

After all, it is possible to combine these two issues to conclude that the indicator function of a word (or a phrase) u is represented by a function that maps a context into a confidence value:

$$\mathcal{X}_u^S : S[u] \to CF(S[u]),$$

where $CF(S[u])$ is measured by the fuzzy version of $\mathcal{D}_{\text{Babel}}$. This provides a way of computing (3–5):

$$CF_l(\mathcal{X}_u, \mathcal{X}_v) \approx 1 - \max\left\{0, \max_{S[\cdot]} \{CF(S[v]) - CF(S[\textit{some } u])\}\right\}. \tag{3–9}$$

Moreover, the general integral on $\mathcal{U}$ in (3–4) w.r.t. μ can be replaced with an integral w.r.t. a weighted counting measure on the space of context, which is now consist of any sequence of words. Hence

$$CF_g(\mathcal{X}_u, \mathcal{X}_v) \approx 1 - \frac{\sum_{S[\cdot]} \mathbb{I}[CF(S[v]) > CF(S[some\ u])] \cdot (CF(S[v]) - CF(S[some\ u]))}{\sum_{S[\cdot]} CF(S[v])}. \tag{3–10}$$

Where $\mathbb{I}[p]$ is a boolean indicator that takes value one when p is true and zero when p is false. It is obvious that (3–9) and (3–10) still keep the local and global property respectively. And they can gracefully degenerate to the non-fuzzy version. But the formulations (3–9) and (3–10) have an important edge over (3–4) and (3–5), since the space $\{S[\cdot]\}$ is no longer intractable as $\mathcal{U}$. In fact, $\{S[\cdot]\}$ is a countable space provided that the collection of words is countable. Therefore the ergodicity in (3–9) and the summation in (3–10) can be done (at least partially) by sampling or ergodic searching.

To summarize, the procedure of judging the confidence of $u \to v$ w.r.t. a corpus $\mathcal{D}$ provided the vocabulary $\mathcal{V}$ is as follows:

1. Generating a collection $\mathcal{S}$ of contexts, $\forall S[\cdot] \in \mathcal{S}$, $S[\cdot]$ is a sequence of words that contains one and only one placeholder $[\cdot]$ and other words from $\mathcal{V}$.
2. Calculating the value of (3–9) or (3–10), during which the ergodicity or summation is done w.r.t. $\mathcal{S}$, i.e., $\max_{S[\cdot]}\{\cdots\}$ is further approximated by $\max_{S[\cdot]\in\mathcal{S}}\{\cdots\}$, and $\sum_{S[\cdot]}\cdots$ by $\sum_{S[\cdot]\in\mathcal{S}}\cdots$.
3. Taking the computed value as the confidence of $u \to v$ respecting (3–7).

The corpus $\mathcal{D}$ should be capable of returning a confidence for any sequence of words. We denote such corpus by $\mathcal{D}_{\text{Oracle}}$ to represent the similarity between our query on the confidence of a sentence and the query on the gradient of a function on given value, which is known as the oracle optimization task. $\mathcal{D}_{\text{Oracle}}$ can be boldly realized by:

$$CF(S) = \mathbb{I}[S \in \mathcal{D}]$$

Where $\mathcal{D}$ is any common corpus. But more properly, we suggest use a web-based search engine as the interactive corpus and taking the number of retrived information as a measure of confidence. This approach provides more flexibility and robustness, and the size of the corpus is no longer a concern.

The algorithm can be formally summarized as Algorithm.3–1:

Algorithm 3–1 Framework of learning hypernym/hyponym relationships from a corpus.

Input: The vocabulary of the language $\mathcal{V}$;
The word *some*;
The interactive corpus $\mathcal{D}_{\text{Oracle}} \in [0,1]^{\{\mathcal{V}^*\}}$;
The compared words u and v, $u, v \in \mathcal{V}$;
The number of context samples $\mathcal{T}$.

Output: The confidence of a hypernym/hyponym pair, $CF(u \to v)$.

$\mathcal{S} = \emptyset$;
for $i = 1$ to $\mathcal{T}$ **do**
  $\mathbf{s} = \mathbf{s}_1 + [\cdot] + \mathbf{s}_2$;
  $\mathcal{S} = \mathcal{S} \cup \mathbf{s}$;
end for
$CF_l = 1 - \max\left\{0, \max_{S[\cdot]\in\mathcal{S}}\left\{\mathcal{D}_{\text{Oracle}}(S[v]) - \mathcal{D}_{\text{Oracle}}(S[\textit{some } u])\right\}\right\}$;
$CF_g = 1 - \frac{\sum_{S[\cdot]\in\mathcal{S}} \mathbb{I}[\mathcal{D}_{\text{Oracle}}(S[v]) > \mathcal{D}_{\text{Oracle}}(S[\textit{some } u])] \cdot (\mathcal{D}_{\text{Oracle}}(S[v]) - \mathcal{D}_{\text{Oracle}}(S[\textit{some } u]))}{\sum_{S[\cdot]\in\mathcal{S}} \mathcal{D}_{\text{Oracle}}(S[v])}$;
return CF_l or CF_g.

Where $\mathbf{s}_i, i = 1, 2$ are sequence of words from $\mathcal{V}$, and the addition in $\mathbf{s} = \mathbf{s}_1 + [\cdot] + \mathbf{s}_2$ means string concat. In practical, the random generated $\mathbf{s}$ is hardly possible a grammatically plausible context, which is a threaten to the performance of the proposed method. Therefore we can substitute selecting contexts from a given corpus $\mathcal{D}$ for random generating, i.e., $\mathbf{s} = S[\cdot]$, where $S \in \mathcal{D}$. The term *some* can be any term in the studied language that means *a subset of*.

3.3 Degeneration of the Proposed Model: I

To see how our proposal can reduce to orthodoxy pattern-based hypernym detection methods, considering an inner-sentence pattern $p[\cdot,\cdot]$ such that if $u \to v$ then $p[u, v]$ is an appropriate sentence/phrase. For example, let:

$$p[\cdot,\cdot] = [\cdot]\ \textit{, and other}[\cdot].$$

If $u \to v$ holds then both following sentences are semantically valid:

v, and other u,

some u, and other u,

Therefore if $p[\cdot,\cdot]$ can successfully address a hypernym/hyponym pair $u \to v$ (whenever $u \to v$, then $p[u, v]$ is a valid sentence, hence might be found in the corpus $\mathcal{D}$) then so can the context:

$$S[\cdot] = [\cdot],\ \textit{and other u}.$$

Hence the pattern-based detection can be fully rehearse from Algorithm3–1 by manully including the pattern information into $\mathcal{S}$. Specifically, given u, v and the pattern $p[\cdot, \cdot]$, we simply add the context

$$S[\cdot] = p[\cdot, u]$$

into $\mathcal{S}$ and the derivation from the pattern-based hypernym/hyponym detection can be completely repeated. In conclusion, our model can gracefully degenerate to state-of-art pattern-based methods, or, to put it in other words, the pattern-based hypernym detection is no more than an instantiation of our model by selecting a special type of contexts as $\mathcal{S}$.

Chapter 4 Applications and Discussions

Having explored the collection of hypernym/hyponym pairs $\mathcal{H}$ from a corpus $\mathcal{D}$, the question as *What is v?* can be readily answered by *v is u* with confidence $CF(u \to v)$ respecting (3–7), (3–9) and (3–10). It should be noted that to **ontologize** the word v (i.e., find the hypernym of v) is a nontrivial task which is still undergoing intensive studies. Among the proposed methods, our formula provides a feasible solution while utilizing few professional knowledge, thus our method is more intelligent in this aspect.

Apart from ontologization, the hypernym/hyponym relationship is helpful in answering other questions. Two representative examples with our proposal applied are presented in this chapter in detail.

4.1 Identification of Synonym

If one word u is the synonym of another word v, then it is necessarily true that the denotations of u and v are the same. In the language of hypernym/hyponym, this is tantamount to the conclude that:

$$((u \to v) \wedge (v \to u)) \textit{ iff } (u \sim v). \tag{4–1}$$

Where $u \sim v$ means that u is a synonym of v, the validity of (4–1) is obvious from the definition of hypernym and synonym. One can naturally translate (4–1) into its fuzzy counterpart so the fuzzy version of hypernym/hyponym relationship can cast its role:

$$CF(u \sim v) = \min\{CF(u \to v), CF(v \to u)\}. \tag{4–2}$$

The r.h.s. of (4–2) can be readily computed from the method proposed in Chapter 3. Since by (4–2), the synonym relationship is proposed together with a level of confidence, the outcome should be more robust and insightful comparing with binary results. Particularly, the synonym judgement in the fuzzy version, (4–2), can always be turned into a binary one by setting a threshold τ, so

$$u \sim v \text{ if } CF(u \sim v) \geq \tau,$$

but not vice versa.

Despite practical aspects as translation, deciphering, and retrieval, identification synonym in an unfamiliar language from the corpus itself is an interesting end to be studied. Addressing the synonym relationship is the elementary task in machine intelligence, which can potentially pave the way for strong artificial intelligence. Our proposal, invoking mainly the mutual substitutability instead of other embedded professional knowledge about the language (which should always be considered as unavailable), is innovative in this sense.

Moreover, we can directly modify the global condition (3–4) to derive a measure of synonym from indicator functions directly, we propose to use:

$$CF(u \sim v) = 1 - \frac{\int |\mathcal{X}_u(e) - X_v(e)| \mathrm{d}\mu}{\int (\mathcal{X}_u(e) + X_v(e))\, \mathrm{d}\mu}, \tag{4–3}$$

whose legitimation depends on the following evidence:

1. $CF(u \sim v) \in [0, 1]$.
2. If $\mathcal{X}_u = \mathcal{X}_v$ a.e. with respect to μ then $CF(u \sim v) = 1$.
3. The form of (4–3) is the combination of

$$CF(p \wedge q) = CF(p) \cdot CF(q)$$

which is another legal generalization of ordinary logic and

$$CF(u \to v) = 1 - \frac{||(\mathcal{X}_v - \mathcal{X}_u)^+||_1}{||\mathcal{X}_v||_1}$$

where $||\cdot||_1$ means one-norm and f^+ means $\max\{f, 0\}$. Now

$$\begin{aligned} CF(u \sim v) &= (1 - \frac{||(\mathcal{X}_v - \mathcal{X}_u)^+||_1}{||\mathcal{X}_v||_1})(1 - \frac{||(\mathcal{X}_v - \mathcal{X}_u)^+||_1}{||\mathcal{X}_u||_1}) \\ &\approx 1 - \frac{||(\mathcal{X}_v - \mathcal{X}_u)^+||_1}{||\mathcal{X}_v||_1} - \frac{||(\mathcal{X}_v - \mathcal{X}_u)^+||_1}{||\mathcal{X}_u||_1} \\ &\approx 1 - a \cdot (||(\mathcal{X}_v - \mathcal{X}_u)^+||_1 + ||(\mathcal{X}_v - \mathcal{X}_u)^+||_1) \\ &= 1 - a \cdot ||(\mathcal{X}_u - \mathcal{X}_v)||_1 \end{aligned} \tag{4–4}$$

This deduction is finalized by let $a = ||\mathcal{X}_u + \mathcal{X}_v||$ to ensure normalization.

Having validated (4–3), we can modify Algorithm 3–1 into an algorithm that judges the confidence of $u \sim v$, the only modification is that we longer compute CF_l or CF_g. Instead the synonym confidence CF_s is computed by:

$$CF_s = 1 - \frac{\sum_{S[\cdot] \in \mathcal{S}} |\mathcal{D}_{\text{Oracle}}(S[u]) - \mathcal{D}_{\text{Oracle}}(S[v])|}{\sum_{S[\cdot] \in \mathcal{S}} \mathcal{D}_{\text{Oracle}}(S[u]) + \mathcal{D}_{\text{Oracle}}(S[v])}. \tag{4–5}$$

Remark: At this point, we can draw a boundary around the semantic tasks that can be done within our framework. Our propsal that consists of the fuzzy model of the universe of discourse and the context-based realization of the computation on such universe can judge any semantic relationship between words u_1, u_2, $\cdots$, u_n such that the relationship can be written as a functional of their indicators:

$$Q(\mathcal{X}_{m_1}(u_1), \mathcal{X}_{m_2}(u_2), \cdots, \mathcal{X}_{m_n}(u_n)).$$

After rewriting Q as a functional of n indicators, its value can be readily evaluated using context sampling as Algorithm 3–1. The hypernym detection and synonym detection are both of this type of semantic tasks with $n = 2$ and the respective functional (3–4) and (4–3).

4.2 Word Embedding

We have proposed that the denotation of a word u is solely determined by the function $\mathcal{X}_u$ which maps $\{\mathcal{V}, [\cdot]\}^*$ into $[0, 1]$, whose value is obtained by querying the oracle corpus $\mathcal{D}_{\text{Oracle}}$. But this is too expensive in time consumption, and insufficient sampling in the space of contexts will lead to a poor result in calculating the confidence of a hypernym/hyponym pair. For example, if $\mathcal{S}$ in Algorithm.3–1 consists of but grammarly

mistaken sentences, then the confidence of $S[v]$ or $S[\textit{some } u]$ will be zero and (3–10) is no longer properly defined.

Therefore one naturally seeks ways of constructing $\mathcal{X}_u$ explicitly from a given corpus $\mathcal{D}$. That is to say, having examined $\mathcal{D}$, one should be able to estimate the confidence of S without asking $\mathcal{D}_{\text{Oracle}}$, where $S \notin \mathcal{D}$. This task has been conclusively studied in branches of studies as recommendation systems, collaborative filtering, and matrix completion (Li, 2016). One of the established methods in recommendation systems suggests that every word or context be encoded as a numerical vector and the fitness of one word u w.r.t. a context $S[\cdot]$, i.e., $CF(S[v])$ is measured by the cosine distance of the vectors representing u and $S[\cdot]$. To learn the vector representation of words, the confidence of the corpus defined as the summation of the confidence of all word-context pairs within the corpus is maximized w.r.t. the vector representations. The vectors are then embedded in the semantic information of words. It should be noted that this formulation can be taken as a direct approximation of our proposal, where a variational approximation is adopted.

The procedure of computing the vector representation of words is summarized as follows, where we have altered the terminologies into the linguistics settings (Levy, 2014)(Goldberg, 2014):

1. Any $w \in \mathcal{V}$ is randomly assigned a vector $\mathbf{w} \in \mathbb{R}^N$.
2. Any context $S[\cdot]$ is assigned a vector $\mathbf{S} \in \mathbb{R}^N$. Specifically,
$$\mathbf{S} = \frac{1}{|S[\cdot]|} \sum_{w \in S[\cdot]} \mathbf{w},$$
where $|S[\cdot]|$ is the number of words (excluded the placeholder) in the context.
3. Take the corpus as the collection of (context, word) pairs:
$$\mathcal{P} = \{(S[\cdot], w) : S[w] \in \mathcal{D}\}.$$
4. Maximize the value :
$$\sum_{(S[\cdot],w)\in\mathcal{P}} \frac{\mathbf{S}^{\mathrm{T}}\mathbf{w}}{\mathbf{S}^{\mathrm{T}}\mathbf{S}}$$
w.r.t. $\{\mathbf{w} : w \in \mathcal{V}\}$ subject to $\forall w \in \mathcal{V}, \mathbf{w}^{\mathrm{T}}\mathbf{w} = 1$. This is usually done by a gradient-based formulation which guarentees a local optimum.
5. The confidence of a new sentence as a pair $(S[\cdot], w)$ is computed by $\frac{\mathbf{S}^{\mathrm{T}}\mathbf{w}}{\mathbf{S}^{\mathrm{T}}\mathbf{S}}$.

This formulation is known as Scenario Vector Decomposition (SVD) (Li, 2016). SVD can be readily absorbed into our framework as a substitution or a variational approximation of $\mathcal{D}_{\text{Oracle}}$.

Despite the combination of SVD and our proposal of fuzzy semantics, studies that aim at interpreting language from a probabilistic perspective attain a similar result (Turian, 2010)(Schnabel, 2015). But in this literature, this formula is named **word embedding**, which is another popular task to be delved into in the nowadays combination between linguistics, statistics and deep learning methodologies (Kusner, 2015). Although word embedding is not taken for granted in our framework, we can still conclude that our model admits updated algorithms as substitutions of its elements and is thence general and inclusive.

4.3 Degeneration of the Proposed Model: II

The word embedding technique that enjoys fame from statistical natural language process community is nothing more than a variational approximation of such mapping. By variational we refer to literature in Bayesian inference as well as quantum mechanics rather than pure maths. In Bayesian inference variational inference means approximating an intractable posterior distribution with a parameterized distribution, while in quantum mechanics it means approximating a base eigenstate (usually the eigenstate of the Hamitonian operator) wave fucntion with a function with fixed form. Word embedding, as we have indicated, tries to obtain the same representation of a word, however, it assumes that the representation is introduced by one specific form: inner product, hence the name reflects itself. Consider the variational approximation below:

$$\hat{CF}(S[u]) = \frac{\mathbf{w}_u^{\mathrm{T}} \mathbf{w}_{S[\cdot]}}{(\mathbf{w}_u^{\mathrm{T}} \mathbf{w}_u)(\mathbf{w}_{S[\cdot]}^{\mathrm{T}} \mathbf{w}_{S[\cdot]})},$$

with

$$\mathbf{w}_{S[\cdot]} = \frac{1}{|S[\cdot]|} \sum_{v \in S[\cdot]} \mathbf{w}_v.$$

And the parameter $\mathbf{W} = \{\mathbf{w}_v\}_{v \in \mathcal{V}}$ needs to be optimized in order to obtain a good approximation:

$$\mathbf{W} = \arg\min_{\mathbf{W}} \left\{ \sum_{(u, S[\cdot]) \in \mathcal{D}} \left[CF(S[u]) - \hat{CF}(S[u]) \right]^2 \right\}.$$

Then the famed word embedding algorithm is recovered.

4.4 Remark on the Essense of Proposed Model

The simple motivation that lies behind our semantic model is to map a word to its confidence in different contexts. This mapping can be traced back to literature about functional analysis where mathematicians tried to establish the correspondence between a vector space $\mathcal{X}$ over a field K and the space of linear functional from $\mathcal{X}$ to K, $\mathcal{X}^*$ (Follang, 2013). For any $x \in \mathcal{X}$, a linear functional can be assigned as:

$$\hat{x}(f) = f(x), f \in \mathcal{X}^*.$$

And the natural introduced mapping:

$$x \to \hat{x}, x \in \mathcal{X}$$

form the prototype of our model. In our discussion the space of word is not necessary linear over a field, but since confidence is normalized to $[0, 1]$, each context can be assigned a function f that maps a word u to the confidence of a sentence with u plugged into the context. And u is featured by nothing more but its value on all contexts, i.e.$\hat{u}$. Therefore our model is nothing more than the combination of fuzzy formulation (where each sentence is assigned a fuzzy confidence instead of absolutely valid or not) and the analog of the natural mapping from a space into its functional dual space.

Chapter 5 Conclusion

In this paper we propose a novel semantic model braced by fuzzy theory. By absorbing fuzzy theory from multiple levels, our model is robust against vagueness and uncertainty in natural language and is analytically convincing. The design philosophy and motivations are split and released in details so the ad hoc aspects of model formulation are avoided to the maximal degree. By invoking the basic semantic properties, the proposed model has several edges over established models:

1. It can trace semantic information between sentences.
2. It can provide numerical results rather than boolean ones, hence is more informative. It naturally adopts web-based, interactive corpus, which is more suitable for modern applications.

After the proposal, we suggest other usages of our model. It should be noted that apart from theoretical utilities, our model is deeply connected with updated statistical tools in the natural language processing community. This connection is insightfully surveyed and discussed. The final remark draws the mathematical essence of our proposal and indicates the nature of the word embedding.

Note that the proposed method is concise in the sense that it is a straightforward combination between fuzzy theory and functional analysis, put into linguistics. But the derivation is still inspiring and profound. Beside semantically plausible results, our work tries to absorb ad hoc techniques as word embeddings into traditional linguistic frameworks to shorten the gap between traditional and modern linguistic studies. And such shortening is precisely the initial motivation behind this work.

Although the modeling hitherto has been self-consistent, there are a few remarks to be recorded for further studies:

1. Can we consider the space of words as a vector space? Be it possible, we can further define a norm on it and continue to reason about other geometrical properties of it such as completeness. However, it seems the linearity of words is not an explicit attribute so we did not reckon so.
2. There should be some metric on the selection of the context set $\mathcal{S}$, e.g. $\mathcal{S}$ should be larger (w.r.t. the order introduced by inclusion) than a minimal set $\mathcal{S}_m$ that **separates points** of $\mathcal{V}$, i.e.

$$\forall u, v \in \mathcal{V}, u \neq v, \exists S[\cdot] \in \mathcal{S}_m, \text{ such that } S[n] \neq S[v].$$

   This is another important property to be addressed if we are to explore more functional propositions of this semantic model.

To conclude, the further formal studies of this semantic model should begin with exerting linearity on the space of words. From which non-trivial properties can be readily deduced analogously as in maths.

Bibliography

1. Rion Snow, Daniel Jurafsky & Andrew Y. Ng. "Learning syntactic patterns for automatic hypernym discovery." *Advances in neural information processing systems.* 2005.
2. Asuka Sumida & Kentaro Torisawa. "Hacking Wikipedia for hyponymy relation acquisition." *Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-II.* 2008.883-888.
3. Sharon A. Caraballo. "Automatic construction of a hypernym-labeled noun hierarchy from text." *Proceedings of the 37th annual meeting of the Association for Computational Linguistics.* 1999. 120-126.
4. Alan Ritter, Stephen Soderland & Oren Etzioni. "What Is This, Anyway: Automatic Hypernym Discovery." *AAAI Spring Symposium: Learning by Reading and Learning to Read.* 2009. 88-93.
5. Roberto Navigli & Paola Velardi. "Learning word-class lattices for definition and hypernym extraction." *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics.* 2010.1318-1327.
6. Erik Sang Tjong Kim. "Extracting hypernym pairs from the web." *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions.* 2007.165-168.
7. Ichiro Yamada, Kentari Torisawa, Junichi Kazama, Kow Kuroda, Masaki Murata, Stijin De Saeger, Farncis Bond & Asuka Sumida. "Hypernym discovery based on distributional similarity and hierarchical structures." *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2-Volume 2. Association for Computational Linguistics.* 2009.929-937.
8. Omer Levy & Yoav Goldberg. "Dependency-based word embeddings." *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers).* Vol. 2. 2014.302-308.
9. Matt J.Kusner, Yu Sun, Nicholas I.Kolkin & Kilian Q.Weinberger. "From word embeddings to document distances." *International Conference on Machine Learning*. 2015.
10. Yoav Goldberg & Omer Levy. "word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method." *arXiv preprint* arXiv:1402.3722.2014.
11. Tobias Schnabel, Igor Labutov David Mimno, & Thorsten Joachims. "Evaluation methods for unsupervised word embeddings." *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. 2015. 298-307.
12. Joseph Turisan, Lev Ratinov & Yoshua Bengio. "Word representations: a simple and general method for semi-supervised learning." *Proceedings of the 48th annual meeting of the association for computational linguistics. Association for Computational Linguistics*.2010.384-394.
13. Gloria Bordogna & Gabriella Pasi. "A fuzzy linguistic approach generalizing boolean information retrieval: A model and its evaluation." *Journal of the American Society for Information Science.* 44.2

1993. 70-82.

14. Francisco Herrera & Luis Martínez. "A 2-tuple fuzzy linguistic representation model for computing with words." *IEEE Transactions on fuzzy systems.* 8.6. 2000. 746-752.
15. Witold Pedrycz & Keun-Chang Kwak. "Linguistic models as a framework of user-centric system modeling." *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans* 36.4 .2006. 727-745.
16. Li Ma. "Clarification on linguistic applications of fuzzy set theory to natural language analysis." *2011 Eighth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD). Vol. 2*. IEEE, 2011.811-815.
17. Richard Khoury, Fakhri Karray, Yu Sun, Mohamed Kemel & Otman Basir. "Semantic understanding of general linguistic items by means of fuzzy set theory." *IEEE Transactions on fuzzy systems.* 15.5 .2007.757-771.
18. Lotfi A. Zadeh. "A fuzzy-set-theoretic interpretation of linguistic hedges." *.Journal of Cybernetics*. 1972.4-34.
19. Lotfi A. Zadeh. "Pruf and its application to inference from fuzzy propositions." *1977 IEEE Conference on Decision and Control including the 16th Symposium on Adaptive Processes and A Special Symposium on Fuzzy Set Theory and Applications.* IEEE, 1977.1359-1360.
20. Irving M. Copi. ”Introduction To Logic. Introduction to logic”. Macmillan, 1978.
21. Gerald B. Folland. ”Real analysis: modern techniques and their applications.” John Wiley Sons, 2013.
22. Zimmermann & Hans-Jürgen. ”Fuzzy set theory—and its applications.” Springer Science Business Media, 2011.
23. Li Xiao-Chun, Li Fang-Qi, Guo Ying & Huang Jin-Chao. "Gaussian Iteration: A Novel Way to Collaborative Filtering." *International Conference on Intelligent Computing. Springer, Cham.* 2016.260-270.

Acknowledgements

I would like to express my gratitude to all those who helped me during the writing of this thesis. I owe my gratitude to my supervisor, Zhexi Wang, who has always been encouraging and dedicating.

Publications

[1] Li Fang-Qi, Ren Xu-Die & Guo Hao-Nan. "Probabilistic Model of Object Detection Based on Convolutional Neural Network." *International Conference in Communications, Signal Processing, and Systems. Springer, Singapore.* 2017.2059-2066.

[2] Li Fang-Qi & Ren Xu-Die. "Laplace Exponential Family Principal Component Analysis" *International Conference in Intelligent Computing. Springer, Singapore*. 2018.322-333.

[3] Li Xiao-Chun, Li Fang-Qi, Guo Ying & Huang Jin-Chao. "Gaussian Iteration: A Novel Way to Collaborative Filtering." *International Conference on Intelligent Computing. Springer, Cham.* 2016.260-270.

[4] Ju Zhen, Lin Xiang, Li Fang-Qi & Wang Shi-Lin. "Lip Segmentation with Muti-scale Features Based on Fully Convolution Network." *2018 IEEE Third International Conference on Data Science in Cyberspace (DSC).* IEEE, 2018.365-370.

[5] Di Chong, Li Sheng-Hong, Li Fang-Qi & Qi Kai-Yue. "A Novel Framework for Learning Automata: A Statistical Hypothesis Testing Approach." *IEEE Access*, 2019.

[6] Li Fang-Qi, Wang Shi-Lin & Liu Gong-Shen. "Bayesian Possibilistic C-Means Clustering Screening for Cervical Cancer" *Information Sciences*, in Press.