# 作者简介及博士学位论文中英文摘要

**论文题目：面向自然语言文本的否定性与不确定性**

**作者简介：邹博伟，男，1984 出生，2011 年 9 月师从于苏州大学朱巧明教授，于 2015 年 12 月获博士学位。**

## 中 文 摘 要


自然语言文本中存在大量否定性与不确定性语言现象，反映了人类在使用语言表达观点时的态度，亦或者语言信息本身的可信度。语言的否定性指，由否定运算符对命题本身或其某一方面的语义进行了反转；语言的不确定性指，包含了情态、言据性、或然性、主观性等任何一类语义，介于肯定和否定语义之间。识别并理解自然语言的否定性与不确定性，对更深层次的自然语言理解具有重要意义，并且随着自然语言处理领域相关应用的不断增长，该研究受到越来越多的关注，如信息抽取、情感分析、信息检索、机器翻译等研究。

面向自然语言文本的否定性与不确定性识别研究主要包含三个子任务：1）触发词检测，即识别出文本中表达否定或不确定语义的关键词；2）覆盖域界定，即在句子内，判定否定或不确定语义的作用范围；3）聚焦点识别，指在覆盖域中识别被否定语义强调的内容。本文研究围绕以上三个任务展开。首先，本文提出了基于树核的覆盖域界定模型，有效并充分地利用结构化句法特征，提高了该任务的性能；其次，本文提出了基于“词-主题”双层结构图模型的聚焦点识别方法，该方法通过上下文信息判断聚焦点；为推动该研究在汉语上的进展，本文构建了首个汉语否定性与不确定性语料库；最后，本文针对汉语的语言特点提出了一套完整的面向汉语的否定性与不确定性识别方法。具体地，本研究的主要内容包括以下四个方面：

1. 基于树核的覆盖域界定模型。覆盖域是指触发词的语义作用范围，大量研究表明，句法特征是判定覆盖域的重要证据，然而，相关工作通常仅考虑平面化的句法特征，即用特征向量来表示句法结构，该表示方法很难恰当并全面地反映触发词与覆盖域在句法结构上的关系。因此，本文提出了两种类型的子树结构来提取触发词与覆盖域之间的关系特征，并利用卷积树核模型衡量这些结构特征之间的相似度，进而确定覆盖域。此外，本文还尝试采用复合核将平面化特征与结构化特征进行融合，提高了现有覆盖域界定方法的性能。

2. 基于“词-主题”双层结构图模型的聚焦点识别方法。不同于面向语音语料的相关研究能够利用重音和语调等特征，面向文本的聚焦点识别研究仅根据词法和句法特征识别聚焦点。通过对聚焦点实例的人工标注及统计，本文发现上下文语境中包含了大量判断聚焦点的线索。因此，本文提出了基于“词-主题”的双层结构图模型，利用上下文中的线索及特征来识别聚焦点。此外，作为无监督模型，该方法还减少了人工标注的开销。实验结果表明，本文的方法能够有效地利用上下文中的信息识别否定聚焦点，其性能优于目前已知最好的系统。

3. 构建汉语否定性与不确定性语料库。目前，面向汉语的否定性与不确定性识别研究进展缓慢，其中最主要的原因是缺乏一个具有一定规模的语料库。因此，本文构建了汉语否定性与不确定性语料库，该语料库是首个已发布的针对文本否定性与不确定性研究的汉语语料库。考虑到语料在领域和文体上应具备异构性，以便充分反映和体现语言现象和特点，汉语否定性与不确定性语料库包含了科技文献、财经文章、酒店评论三个类别，其规模达到 16,841 句，包含 6,429 个实例，与目前英文中使用最频繁的 BioScope 语料库规模相近。相关统计和实验结果表明，本文构建的语料库较全面地体现了汉语中否定性与不确定性语义的特点，为相

关研究提供了语料资源支持。

4. 面向汉语的否定性与不确定性识别研究。由于汉语与英语在语法结构及语义表达等诸多方面均存在较大差别，直接将英语中的否定性与不确定性识别方法应用在汉语上时，系统性能大幅下降。因此，针对触发词检测，本文提出了一套适用于汉语的新特征，尤其是词素特征，同时还采用了跨语言触发词扩展策略，识别出现频率较低的触发词；针对覆盖域界定，本文提出了基于元决策树的方法，该方法有效融合了序列化特征和结构化特征。本文工作构建了首个面向汉语的否定性与不确定性识别系统，希望能够为相关研究提供基线系统，并促进该研究在汉语上的开展。

总之，本文致力于面向自然语言文本的否定性与不确定性识别研究，一方面提出了有效方法来提高相关任务的性能，一方面尝试推动该研究在汉语上的进展。期待本文取得的初步成果能够对该领域的相关研究产生一定的参考价值，促进自然语言深层
理解技术的发展。

**关键词：否定，不确定，触发词检测，覆盖域界定，聚焦点识别**

# Research on Negation and Uncertainty Identification on Natural Language Text

**ZOU Bowei**

## ABSTRACT

Negation and uncertainty, very common phenomena in natural language, reflect either the attitudes of human beings on expressing views in natural language or the credibility of linguistic information. While negation is a grammatical category which comprises various kinds of devices to reverse the truth value of a proposition, uncertainty is a grammatical category which expresses a statement in terms of degree of modality, evidentiality, probability, and subjectivity. Obviously, in recent years, negation and uncertainty identification plays a critical role in deep natural language understanding and has being drawn more and more attentions with increasing applications in related topics, such as Information Extraction (IE), Sentiment Analysis (SA), Information Retrieval (IR), Machine Translation (MT).

The research of negation and uncertainty identification on natural language text contains three main sub-tasks. The first is cue detection, which aims at detecting whether there is a negative or uncertain keyword in the given text. The second is scope resolution which aims to determine the linguistic scope of a given cue in sentence. The third is focus identification, which aims at identifying the most prominent or explicit part negated by a negative cue in scope. In this paper, firstly, we propose a tree kernel-based model for scope resolution, which utilizes the structured syntactic features effectively and improves the performance of scope resolution. Then, we propose a "word-topic" graph model which identifies focus with context. To promote the advance of related

research on Chinese, we construct a Chinese negation and uncertainty corpus. Finally, we propose effective methods according to the characteristics of Chinese language. The main contents of our research work can be summarized as follows:

1. Proposing a tree kernel based negation and uncertainty scope resolution model. As the scope is defined as the semantic scope of a cue, the syntactic features are always employed as the important evidence to determine scope. However, the related work only uses the flat syntactic features which are generally represented by a feature vector. It is difficult to felicitously and fully reflect the characteristic on syntactic structure. Therefore, we propose two kinds of sub-trees related to cues and measure the similarity of these structures by a convolution tree kernel. Moreover, we also fuse both the flat features and the structure features by a composite kernel, which improves the performance of scope resolution.

2. Proposing a "word-topic" structured bilayer graph model based focus resolution method. Different from the focus identification on speech corpora which contains more stress or intonation information, the research on text corpora only utilize the morphological or syntactic features to identify negation focus. We find that the contextual discourse information plays a critical role on focus identification, which is determined by the semantic relatedness between the negation expression and the emphasis of author in context. On the basis, we propose a "word-topic" structured bilayer graph model to evaluate the effect of the contextual discourse information on the negation focus. Moreover, as an unsupervised method, it also reduces the time-consuming manual annotation for negation focus. The experimental results show that this method is effective for negation focus identification and outperforms the state-of-the-art negation focus identification systems.

3. Constructing the Chinese negation and uncertainty corpus. Currently, the scarcity of linguistic resource seriously limits the advance of the research of negation and uncertainty identification on Chinese. Therefore, we construct the Chinese Negation and Uncertainty Corpus (CNeUn) which is the first and the only Chinese corpus for this research as far as we know. Considering the heterogeneity and characteristics of language use in different domain and literary style, the CNeUn corpus consists of three different sources and types, including scientific literature, product reviews, and financial articles. It contains 16,841 of sentences and 6,429 of instances, similar to the scale of BioScope which is the most frequently used corpus in English. The statistics and the experiment results show that the CNeUn corpus can adequately reflect the linguistic characteristics of negation and uncertainty in Chinese and provide the support for related research.

4. Proposing negation and uncertainty identification methods in Chinese. Since the great difference between English and Chinese on grammatical structure and semantic expression, the performance of the state of the art method in English is low when it is used on Chinese corpus directly. Therefore, for cue detection, we propose a feature-based sequence labeling model with series of features, especially the morpheme feature. In addition, a cross-lingual cue expansion strategy is proposed to increase the coverage. For scope resolution, we propose a meta-decision tree model to integrate serialized features and structured features. As far as we know, this work is the first research which explores on Chinese negation and uncertainty identification systematically.

In conclusion, this paper focuses on the negation and uncertainty identification on natural language text. On the one hand, we propose effective methods to improve the performance of the negation and uncertainty identification. On the other hand, we also try to promote the research progress in Chinese language. This research has achieved some preliminary results, which we hope can not only be helpful to other researchers in this area but also promote the development of deep natural language understanding.

**Key words: Negation, Uncertainty, Cue detection, Scope resolution, Focus identification**