**基于自然语言的语义分析在金融数据分析系统中的应用开题综述**

1. 研究背景及意义

大数据时代的到来，让人们认识到从数以TB甚至PB计的数据中，提取出对于用户有用的信息至关重要。与此同时在金融领域，每天产生的交易信息数据也是非常之庞大的，对这些数据的合理分析利用自然也成为金融业更好发展的辅助。这其中针对不同用户的个性化分析需求颇为重要，海量分析自然会使用计算机，但是人类的自然语言较之机器语言存在“重复性”，“二义性”。如何建立自然语言和分析系统之间的沟通桥梁成为一大难题。

当今社会日益信息化，人们越来越希望用自然语言(即人们日常使用的各种通俗语言)同计算机交流。自然语言理解是计算机科学中的一个引人入胜的、富有挑战性的课题。从计算机科学特别是从人工智能的观点看，自然语言理解的任务是建立一种计算机模型，这种计机模型能够给出象人那样理解、分析并回答自然语言的结果。

自然语言的语义是指信息包含的概念和意义。语义不仅表述事物本质，还表述事物之间的因果、上下位、施事等各种逻辑关系。因此，语义是对事物的描述和逻辑表示。语义分析就是对信息所包含的语义的识别，并建立一种计算模型，使其能够像人那样理解自然语言。语义分析是自然语言理解的根本问题，它在自然语言处理、信息检索、信息过滤、信息分类、语义挖掘等领域有着广泛的应用。在互联网时代，面对海量的信息资源，要想准确地进行信息抽取，检索所需信息、挖掘潜在的信息价值、提供智能的知识服务，都离不开面向机器理解的语义分析。尤其在大数据环境下，语义分析的地位越来越凸显出来。

本课题主要的研究工作是以实习公司的金融股票分析系统为背景，将基于自然语言的语义分析运用到该系统中去。针对上文总结的自然语言较之机器语言存在的“重复性”“二义性”，以及用户难以描述真正需求等难题，为用户和分析系统提供自然语言处理的接口，方便用户迅速表达自己想要的分析需求，为系统提供更加人性化的用户体验，具有很强的现实意义。

1. 国内外研究现状

近年来，国内的金融市场呈现着一片大好局面，在股票领域，越来越多公司的创办使得股票种类越来越多。随之每日交易产生的数据也是呈指数级别的增长。目前市场上有万得和同花顺公司提供这些金融数据的服务，即便如此它们目前也仅仅停留在原数据的展现上，或者对数据进行基本层面上的统计处理。还没有一个专门针对金融数据做分析的软件，当然将自然语言处理应用也是少有在金融领域的应用。

与此同时自从自然语言理解这一学科在20世纪40年代末期诞生以来，自然语言理解在各个独立的领域内取得了巨大的成功，人们逐渐认识到在领域内实现自然语言理解是可行的。20世纪70年代初期，对语言理解对话系统的研究取得进展，伍兹(Woodz)的LUNER系统、威诺甘德(Winogand)的SHRDLU系统和山克(Sehank)的MARGIE系统都是语言理解对话系统的典型实例；20世纪80年代，机器学习研究又十分活跃，并出现了许多具有较高水平的实用化系统。其中比较著名的有美国的METAL和LOGOS，日本的PⅣOT和HICAT，法国的ARIANE以及德国的SUSY等系统，这些系统是自然语言理解研究的重要成果，表明自然语言理解在理论和应用上取得了突破性成果。

纵观这些成功的自然语言理解系统，都是将应用范围确定为某一特定的领域，具有很强的针对性：而且这些系统基本上都采用基于知识的方法建立，在领域内的应用能够达到专家级水平。随着专家系统和知识工程等技术的日益成熟，以及人们对自然语言理解研究的逐渐深入，建立某一领域内的基于知识的自然语言理解系统是能够取得成功的、是可行的。

由此看来，将自然语言的语义分析应用到金融股票这个领域是可行的，但又是非常具有价值的。因此针对用户在金融分析需求上，自然语言与机器语言难以沟通的实际问题，开发一套应用于金融领域的自然语言处理系统，提供用户分析需求对应的自然语言分析功能，具有很强的显示意义和市场前景。

1. 相关技术分析

3.1 分词算法

分词算法是自然语言处理的第一步，我国对分词与词性标注的研究已经有20多年的历史。围绕着基于规则的理性主义方法和基于统计的经验主义方法，提出过许多卓有成效的理论与方法。

经过20余年的发展，分词与词性标注的处理已经达到较高水平。中科院计算技术研究所张华平、刘群研制的ICTCLAS中文词语分析一体化系统是我国目前最先进的分词与词性标注软件之一，在2003年SIGHAN分词测评中，它参加了几乎所有的测试．都取得了较好的成绩，并有多项测试成绩排名第一。该系统采用隐马尔科夫模型．建立切分词图。词语粗分阶段，先得出N个概率最大的切分结果。然后利用角色标注方法识别未登录词。并计算其概率，将未登录词加入到切分词图中，之后视其为普通词处理。最终进行动态规划，优选出N个最大概率切分标注结果。利用N-最短路径方法进行词语粗分．可以较好地解决词语切分问题。同时将未登录词与歧义问题保留到下一个过程。基于角色标注的未登录词识别方法，可以克服候选词语选取的盲目性，并能计算出候选词的概率。基于隐马尔科夫的一体化方法能够将未登录词、歧义消除与普通词的处理统一起来进行，最终得到满意的识别效果。该系统作为开放软件，已经将完备的文档连同源代码上传到中文自然语言处理开放平台上，免费供大家研究与使用。

其实目前用python语言写的比较出名的是结巴分词，

它的特点是：

1. 支持三种分词模式：

a.精确模式，试图将句子最精确地切开，适合文本分析；

b.全模式，把句子中所有的可以成词的词语都扫描出来, 速度非常快，但是不能解决歧义；

c.搜索引擎模式，在精确模式的基础上，对长词再次切分，提高召回率，适合用于搜索引擎分词。

1. 支持繁体分词
2. 支持自定义词典

3.2 语义分析（简单句、复杂句处理）

所谓语义分析，指的是根据句子的句法结构和句中每个实词的词义推导出能够反映这个句子意义(即句义)的某种形式化表示。例如对于句子:“张三吃了苹果”和“苹果被张三吃了”，虽然它们的表述形式不同，但表示成语义的形式就统一为:“吃(张三，苹果)”。

对句子进行正确的语义分析，一直是从事自然语言理解研究的学者们追求的主要目标。然而，经过几十年的发展，目前还没有出现太多使用学习的方法来获取详细语义理解知识的研究。基于语料库的词义消歧研究虽然触及了语义问题，但是这只是在理解单个词的层次上，而不是进行整个句子的理解.关于信息抽取的研究也触及了一些语义理解，然而己有系统只是使用相当低水平句法模式分析等方法来获取特殊的目标短语。

Gildea 等人使用经验主义的方法进行语义角色标注(SemanticR oleL abeling)的研究可谓是对语义理解的一种新的探索。这一研究领域又被称为“浅层语义分析(Shallow Semantic Parsing)". 所谓浅层分析，即该方法并不对整个句子进行详细的语义分析，而只是标注句子中的一些成份为给定动词的语义角色，这些成分作为此动词的框架的一部分被赋予一定的语义含义，例如“[委员会Agent][明天Tmp]将要[通过V[此议案Passive].”其中，“通过”为目标动词，“委员会”、“此议案”和“明天”分别是其施事、受事和发生的时间。另外，浅层语义分析不考虑时态信息，例如“他将来北京。”与“他来北京了。”，虽然时态并不相同，但是浅层语义表示是相同的，同为:“来(他，北京)”。同时，浅层句法分析也不考虑目标动词改变但语义不变的情况，例如“他出生于1969年3月18日。”与“他的生日是1969年3月18日。”，虽然它们的语义相同，但是浅层句法分析的表示结果并不一样，需要根据具体的应用进行更深入的处理。

进行浅层语义分析的基础技术，如词性标注、句法分析、统计学习等目前己经比较成熟。同时浅层语义分析在问答系统、信息抽取、机器翻译等领域有着广泛的应用。

3.3建立领域语料对应于机器语言的语料库

语料库(corpora)是应用计算机技术对海量自然语言材料进行处理(包括预处理、语法自动附码、自动句法分析、语义分析等)、存储, 以供自动检索(retrieval)、索引(concordance)以及统计分析的大型资料库。它是按照明确的设计标准为某一具体目标而建立的语言资料库(Atkinsand Clear 1992 :5 , 引自Granger 1996)。如果作进一步区分, 还可以把语料库与大型文档资源库(text archives)区别开来。前者具有明确语料选择比例和设计标准, 而后者更注重语料容量和语料来源的多样性(Edwards 1993)。所谓自然语言是指任何人类在童年习得的语言(Leech 1987 :1)。使用这一概念的意义在于, 把人类自然语言同人工编制的人工语言(artificial languages)(如C 语言、BASIC 、JAVA 、FORTRAN 等)区别开来;在对语言描述和研究中, 自然出现的语料不同于研究者为某一目的而引出的材料(elicited data)。语料库的研究对象是人类语言自然运用(performance), 而不是内在语言能力(competence)。语料库的应用主要在于以下几个方面:1)自然语言处理(包括自动语言识别、自动语法附码、句法分析、语义分析、知识表达、机器翻译等)。基于语料库方法可以提供有关语言结构和特征的可靠信息, 而这些信息是内省法研究或心理测验获取的信息所不能比拟的。语料库是人们广泛深入理解语言所必需的工具(Edwards 1993 , Leech 1991 , 1992 , Svartvik 1992)。2)词典编纂。语料库能够提供更为完备和详尽的关于词义和词汇搭配信息(Kjellmer 1984 , Sinclair 1982 , 引自Edwards 1993)。3)语言学习与语言教学。早在本世纪60 年代, 美国Brown 大学创建了最早的语料库BROWN CORPUS , 容量为100万词, 广泛搜集了当时美国英语各种文体的语言材料。与此相呼应的是70 年代初在欧洲同样容量的LOB 语料库, 搜集了同年代的英国英语材料。基于这两个语料库, 开展了大量的重要研究, 研究的重点是对英语语法的描述和分析(Rundell 1996)。另外, 70 年代初还建成了50 万词的英语口语语料库London-Lund Corpus(LLC)。但对于大规模的词汇研究(如词典编纂)而言, 100 万词的容量显然不能满足要求。80 年代初, 由Birmingham 大学John Sinclair 主持建成了专门用于词典编纂的大型语料库BCET(Birmingham Collection of English Texts)。该语料库最初容量为730 万词, 到80 年代中期, 其迅速扩充为2000 万词。与此同时, 朗曼公司词典部开始建立自己的语料库Longman-Lancaster 语料库, 容量达到3000 万词。在我国, 上海交通大学杨惠中教授在80 年代末主持建成了容量为300 万词的科技英语语料库(JDEST , 1997 年扩充至360万词), 主要用于大学英语教学大纲制定和教材开发。80 年代建成的语料库还有Lancaster 的英语口语语料库(ESC , 1984 —1987)、用于跨文化比较的PIXI 语料库。90年代初, 包含英国书面英语和口语容量为1 亿词的英国国家语料库建成(BNC , British National Corpus , 1994), 而伯明翰大学的BCET 语料库进一步扩展为英语库(Bank of English), 容词量为两亿。到90 年代中期, 朗曼公司在BNC 库中又增加1 亿词美国英语, COBUILD 的Bank of English 进而增长至3 .2亿词(Rundell 1996)。

1. 研究目标及内容

4.1论文研究目标

本文的研究工作以实习公司的金融股票数据分析项目为背景，该项目是将大数据时代的数据处理分析新技术用于金融领域，为海量金融数据处理提供服务。减轻金融工作者的压力，为金融分析提供量化的数值图标展现。本文目标是将自然语言的语义分析处理应用到该系统中，为用户提供一个分析需求自然语言与分析系统之间的桥梁，让用户的需求能更加快速准确地被机器语言理解，最终给出分析报表。

4.2 论文研究内容

针对上述目标，本文研究内容包括：

1、系统需求分析：以实习公司的金融股票数据分析系统为背景，将自然语言的语义分析在金融股票领域应用系统涉及的业务实体、业务流程、功能和非功能性需求进行分析。

2、相关技术研究，本系统所涉及的相关技术主要有：

1. 分词算法实现：进行词的划分、此类划分和词性标注，识别出汉语、数字、英文标识符、标点符号等，用来实现自然语言到语言理解系统的词法状态转换。
2. 简单句处理：主要包括名词快处理、动词块处理、事件处理以及事件关系处理等
3. 复杂句处理：复杂句处理主要目的是理解较复杂的复合句，理解结果是两个或多个动态事件。这些动态事件前后是按照一定的逻辑关联如条件、原因、结果等组合在一起来说明一个主题。对整个复杂句的理解是用循环加递归实现的，最终的分析结果会是一个动态事件集。
4. 建立领域语料对应于机器语言的知识库：语义分析后产生的不同结果事件、关系和对象事实等，根据金融领域的常识知识规则和领域内知识，建立整个事实对象之间的关联，并将金融领域问题语义分析结果转换为计算机领域的特定形式。
5. 机器学习算法:该算法是一类从数据中自动分析获得规律，并利用规律对未知数据进行预测的算法。因为学习算法中涉及了大量的统计学理论，机器学习与统计推断学联系尤为密切，也被称为统计学习理论。

3、系统的设计与分析，包括：系统结构设计，词法句法分析模块，语义分析模块，查找知识库模块。

4、系统测试：筛选实习公司该项目上线后服务的1000名客户对象，收集客户所要表达的需求和该自然语言处理系统所提取的分析结果。计算出正确率，本系统要求正确率在80%以上

5、系统应用：本文将对本应用在金融股票数据分析项目上线运行3个月的应用效果进行分析和总结。

1. 技术路线及创新意义

本课题的关键问题主要有：

1. 自然语言存在“重复性”问题：提供词法句法分析，将句子拆分成具有独立意义的词组。
2. 分词后存在“二义性”问题：结合静态知识库、推理规则以及推理过程中产生的动态知识库，进行逐词逐句的分析理解，消除词语和句子的歧义，生成自动建模所需要的全部事实:关系类、事件类、对象类等。
3. 中间分析结果无法被机器识别：建立金融股票领域知识规则对应于机器语言的关联，最终将数据分析需求理解结果转换为机器能理解的形式。
4. 知识库的固定不变问题：使用机器学习，通过不断地扩充语料样本集来扩充知识库。

本课题的创新意义主要有：

本课题基于公司的业务需求开展，与实际联系紧密，体现了很好的应用创新性。将自然语言处理运用到金融领域，实现自然语言的分析需求与分析工具方便快捷地交互。方便用户使用分析系统，有利于分析系统更加快速的将用户的需求转化为机器能够理解的操作原语句，提高分析系统的用户体验。

1. 参考文献
2. Chen K J, Liu S H. Word identification for Mandarin Chinese sentences[C]//Proceedings of the 14th conference on Computational linguistics-Volume 1. Association for Computational Linguistics, 1992:101-107.
3. Xue N. Chinese word segmentation as character tagging[J]. Computational Linguistics and Chinese Language Processing, 2003, 8(1): 29-48.
4. Peng F, Feng F, McCallum A. Chinese segmentation and new word detection using conditional random field [C]//Proceedings of the 20th international conference on Computational Linguistics. Association for Computational Linguistics, 2004: 562.
5. Sproat, Richard, Chilin Shih. A statistical method for finding word boundaries in Chinese text[J]. Computer Processing of Chinese and Oriental Languages,1990(4):336–351
6. Ge X, Pratt W, Smyth P. Discovering Chinese words from unsegmented text (poster abstract)[C]//Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 1999: 271-272.
7. Peng F, Schuurmans D. Self-supervised Chinese word segmentation[M]//Advances in Intelligent Data Analysis. Springer Berlin Heidelberg, 2001: 238-247.
8. Peng F, Schuurmans D. A Hierarchical EM Approach to Word Segmentation[C]//NLPRS. 2001: 475-480.
9. Feng H, Chen K, Kit C, et al. Unsupervised segmentation of Chinese corpus using accessor variety[M]//Natural Language Processing–IJCNLP 2004.Springer Berlin Heidelberg, 2005: 694-703.
10. Jin Z, Tanaka-Ishii K. Unsupervised segmentation of Chinese text by use of branching entropy[C]//Proceedings of the COLING/ACL on Main conference poster sessions. Association for Computational Linguistics, 2006: 428-435.
11. Chen S, Xu Y, Chang H. A Simple and Effective Unsupervised Word Segmentation Approach[C]//AAAI. 2011.
12. Wang H, Zhu J, Tang S, et al. A new unsupervised approach to word segmentation[J]. Computational Linguistics, 2011, 37(3): 421-454.
13. Magistry P, Sagot B. Unsupervized word segmentation: the case for mandarin chinese[C]//Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2. Association for Computational Linguistics, 2012: 383-387.
14. Grishman R, Sundheim B. Message Understanding Conference-6: A Brief History[C]//COLING. 1996, 96: 466-471.
15. Yarowsky D. Unsupervised word sense disambiguation rivaling supervised methods[C]//Proceedings of the 33rd annual meeting on Association for Computational Linguistics. Association for Computational Linguistics, 1995:189-196.
16. Agichtein E, Gravano L. Snowball: Extracting relations from large plain-text collections[C]//Proceedings of the fifth ACM conference on Digital libraries. ACM, 2000: 85-94.
17. McClosky D, Charniak E, Johnson M. Effective self-training for parsing[C]//Proceedings of the main conference on human language technology conference of the North American Chapter of the Association of Computational Linguistics. Association for Computational Linguistics, 2006:152-159.
18. 翟凤文, 赫枫龄, 左万利. 字典与统计相结合的中文分词方法[J]. 小型微型计算机系统, 2006, 27(9): 1766-1771.
19. 王伟, 钟义信. 一种基于 EM 非监督训练的自组织分词歧义解决方案[J].中文信息学报, 2001, 15(2): 38-44.
20. 王宁, 黄锦辉. 中文金融新闻中公司名的识别[J]. 中文信息学报, 2002,16(2): 1-6.
21. 赵健. 条件概率模型研究及其在中文名实体识别中的应用[D]. 哈尔滨工业大学博士论文, 2006：1757-1862
22. 姚天顺;朱靖波;张琍自然语言理解--一种让机器懂得人类语言的研究 2002
23. 赵克;许威一个基于网络服务的平面几何辅导专家系统[期刊论文]-计算机应用研究 2003(6)
24. 于会;李伟华;陈栋专家系统中的知识表示及其实时处理方法研究[期刊论文]-微电子学与计算机 2005(05)
25. 吴中兴;赵克;胡钢伟;许威概念从属树-一种新的树模型设计[期刊论文]-计算机应用 2004(06)
26. 赵克;面向并行工程的设计过程理论与实现方法研[学位论文] 1995
27. 亿珍珍;面向产品设计的自然语言理解语义分析研究[学位论文] 2005