基于词频和语义信息的组合型歧义消解*

丁德鑫[1,3] 曲维光[1,3] 于丽丽[2] 陈小荷[2] 李惠[2]

1. 南京师范大学计算机科学与技术学院 南京 210097 2. 南京师范大学文学院 南京 210097

3. 江苏省信息安全保密技术工程研究中心 南京 210097

E-mail: dingdexin2@126.com

摘 要：组合型歧义切分是汉语自动分词的难点之一。本文挖掘歧义字段上下文的相对词频信息和语义信息，建立语境计算模型。首先基于相对词频比，建立 RFR_SUM 模型，其次采用类似 K 近邻的分类思想，利用知网，建立语义相似度计算模型，最后尝试两个模型的结合，进行歧义消解。以 1998 年半年《人民日报》作为实验语料，对 20 个典型的组合歧义字段进行消歧，两个模型及二者结合构成模型的 F 值分别为 96.01%、93.42%、96.52%，高于常用的条件随机场和最大熵模型，取得了良好的效果。

关键词：中文自动分词，组合歧义，语境信息，RFR_SUM，语义计算

Resolving Combinational Ambiguity Based on Word Frequency and Semantic Information

DING Dexin[1,3] QU Weiguang[1,3] YU Lili[2] CHEN Xiaohe[2] Li Hui[2]

1. School of Computer Science and Technology, Nanjing Normal University, Nanjing 210097;

2. School of Chinese Language and Literature, Nanjing Normal University, Nanjing 210097;

3. The Research Center of Information Security and Confidentiality Technology of Jiangsu Province, Nanjing 210097

E-mail: dingdexin2@126.com

Abstract: Combinational ambiguity is one of the vital issues in Chinese word segmentation. By mining relative frequency information and semantic information of context, we establish the context calculation model. Firstly create the RFR_SUM model, which is based on the sum of relative frequency ratio of words in context, and then use HowNet, the classification method similar to the k-nearest neighbor to calculate semantic similarity respectively. And finally attempt to combine the two models. 20 typical combinational ambiguity words are tested by using half year corpus of the 1998 "People's Daily", and the average F-score of the methods are 96.01%,93.42%,96.52%,better than the often-used CRF and Maximum entropy model.

Keywords: Chinese word segmentation, Combinational ambiguity, RFR_SUM, Semantic calculation

1 引言

汉语自动分词是自然语言处理的基础，直接影响到后续深层的语言分析和理解[1]。歧义切分又是影响分词系统切分精度的重要因素[2]。歧义切分有两种基本类型：交集型歧义和组合型歧义。迄今为止，汉语分词歧义的研究多集中在交集型歧义，针对组合型歧义的则较少。大多数交集型歧义通常可根据字段内部的信息[2]，或以句法为主的局部上下文信息[3]予以解决。组合型歧义

*基金项目：国家自然科学基金项目（60773173）；国家 973 项目（2004CB318102）；国家社科基金项目(07BYY050)；江苏省社科基金项目（06JSBYY001）.

的处理策略则不相同，往往须诉诸更多的上下文信息(其中语义信息起着十分重要的作用)[4]。本文仅讨论组合型歧义切分字段的消解策略。

2 消解模型

2.1 RFR_SUM 分类模型

曲维光提出了相对词频比的概念，据此建立语境计算模型，利用歧义字段前后语境信息对组合型分词歧义进行消解。该模型不仅考虑了语境中存在的词频，而且考虑了语境中出现词语相对于整个语料词频的相对比率，用相对词频比来模拟人们判断语境中出现词语对消歧的重要程度；同时又区分了语境的位置，将语境分为前语境和后语境，从而把前后语境出现的词语区分开来，提高了语境信息计算的准确性。具体分类算法详见文献[5,6]。

2.2 基于知网的语义相似度计算消歧模型

鉴于RFR_SUM模型考虑了上下文的词频，而组合型歧义的消解往往须诉诸更大范围的上下文信息，其中语义信息起着十分重要的作用[4,7]。所以我们尝试利用知网计算语义相似度，通过计算待消歧句子与两类训练样本句子集合的相似度的方法来解决组合型歧义。计算模型分为三个层次：词语与词语相似度计算、句子与句子相似度计算、歧义消解。

1 词语相似度的计算

利用HowNet(知网)[8]进行词语相似度的计算，学者已经进行了诸多研究。本文用到的词语相似度部分的计算借鉴文献[9]的词语相似度研究。根据刘群等提出的《知网》词汇语义相似度计算方法，两个孤立词汇之间的相似度问题可以归结到两个概念之间的相似度问题，假定两个汉语词汇W_1和W_2，如果W_1有n个义项(概念)：$S_{11}, S_{12}, \ldots, S_{1n}$；W2有m个义项(概念)：$S_{21}, S_{22}, \ldots S_{2m}$，则$W_1$和$W_2$的相似度是各个概念的相似度之最大值，即：

$$Sim(W_1, W_2) = \max_{i=1\ldots n, j=1\ldots m}(sim(S_1, S_2))$$

概念用义原来表示，义原的相似度计算是概念相似度计算的基础。义原相似度是根据义原所处的树状层次体系中语义距离计算得到：

$$Sim(p_1, p_2) = \lambda / d + \lambda$$

其中P_1和P_2表示两个义原，d 是P_1和P_2在义原层次体系总的路径长度，为一个正整数，λ为可调节参数。具体算法详见文献[9]。

2 句子相似度的计算

对于包含歧义字段的2个句子：Sen1,Sen2，我们分别抽取出Sen1，Sen2的上下文一定窗口内(本文设定窗口为5)的词构成四个集合：frontsen1，backsen1，frontsen2，backsen2。frontsen1与frontsen2、backsen1与backsen2的相似度的计算过程如下：

令SIZE=max{|Set1|,|Set2|}，|Set1|和|Set2|分别表示2个集合当前拥有的词的数量，即

score=0.0;

while (|Set1|>0 or |Set2|>0)

{

求出2个集合所有组合中相似度最大的一对词$W_i \in Set1$和$W_j \in Set2$；

score = score + Sim (W_i, Wj);

Set1 = Set1−{W_i};

Set2 = Set2−{W_j};

}

其中Sim (W_i, Wj)计算两个词的词语相似度，则上下文的相似度：

SimContext (Set1, Set2) = score/SIZE。

句子的相似度为歧义字段上下文的相似度之和：

SimSen (Sen1, Sen2)=SimContext(frontsen1, frontsen2)+SimContext (backsen1, backsen2)。

3 歧义消解

有了上述句子与句子的语义相似度计算，我们就可以对句子中待消歧字段做出从分从合的判定，具体采取类似机器学习中K近邻分类的方法。设trainsenset1表示从分的训练句子集合，trainsenset2 表示从合的训练句子集合，testsen为待测试的句子。判定算法如下：

Sum1=0, Sum2=0;

m=|trainsenset1|, n=|trainsenset2|, 即分别为两类训练样本的句子数。

For (i=0; i<m; i++)

计算句子的相似度: SimSen (trainsenset1(i),testsen)，并存入double数组SimSet1s。

Sum1+=SimSet1s中的KnnNum个最大值。

Sum1/=KnnNum。

For (j=0; j<n; j++)

计算句子的相似度: SimSen (trainsenset2(j),testsen)，并存入double数组SimSet2s。

Sum2+=SimSet2s中的KnnNum个最大值。

Sum2/=KnnNum。

若 Sum1>Sum2,歧义字段从分；

若 Sum1<Sum2,歧义字段从合；

若 Sum1=Sum2,歧义字段按强势切分。

2.3 RFR_SUM 与语义相似度模型的结合

基于相对词频的 RFR_SUM 模型考虑了上下文的词频的信息，基于语义相似度的模型考虑了上下文的语义相似信息，各有特点，把二者结合起来，可以实现优势互补。结合的方法如下：

令 RfrSum1 和 RfrSum2 分别是 RFR_SUM 的从分从合的值，SimSum1 和 SimSum2 分别是语义相似度模型的从分从合的值，新的判定公式如下：

SUM1= RfrSum1 * SimSum1;

SUM2= RfrSum2 * SimSum2;

若 SUM1 > SUM2,歧义字段从分；

若 SUM1 < SUM2,歧义字段从合；

若 SUM1 = SUM2,歧义字段按强势切分。

3 实验及其分析

3.1 实验数据

本文使用 1998 年上半年《人民日报》的标准语料做实验语料，共计 1300 万字。对消歧结果进行评测的指标主要是正确率、召回率以及 F 值，计算公式分别如：

正确率(P)=正确判定个数/判定的总数量；

召回率(R)=正确判定个数/测试总数；

F 值=2*P*R/(P+R)。

我们对诸多文献中用过的 20 个典型常用组合型歧义字段[10]进行实验，抽取包含歧义字段的句子，由于一些词的从分或从合相应句子数量过少，我们适当增加了一些例句。实验时，划出 70%的例句作为训练，余下的 30%作为开放测试，具体的从分从合句数如表 1(Num_co、Num_se 分别代表从合的例句数和从分的例句数)。

字段	Num_co	Num_se	Baseline	字段	Num_co	Num_se	Baseline
才能	108	766	87.45	一块	30	247	89.29
都会	31	379	91.94	一起	1406	206	87.19
个人	213	1860	89.71	正当	176	76	69.74
将来	254	30	89.53	中共	730	26	96.48
马上	231	30	88.61	中学	323	34	89.81
人才	346	34	90.43	中将	135	131	50.62
上来	124	233	64.81	走向	68	709	91.03
现在	378	20	95.00	一定	704	9	98.60
可以	244	17	92.50	总会	117	52	69.23
学会	253	183	58.02	中长期	164	20	89.29

表 1 实验例句统计

Baseline 指的是，测试样本都标成类别个数多的样本所在的类时的正确率，其中这些词的平均 Baseline 为 83.96%。

3.2 实验数据及分析

我们分别利用 RFR_SUM、语义相似度计算模型、两者结合进行了测试，结果如表 2：

字段	RFR_SUM	语义计算	两者结合	字段	RFR_SUM	语义计算	两者结合
才能	95.44	93.92	96.96	一块	92.86	89.29	94.05
都会	98.39	95.16	98.39	一起	97.93	94.21	98.14
个人	97.27	92.44	97.43	正当	94.74	96.05	96.05
将来	97.67	93.02	97.67	中共	99.56	97.80	99.56
马上	96.20	98.73	97.47	中学	99.07	88.89	99.07
人才	97.39	89.57	98.26	中将	92.59	95.06	93.83
上来	94.44	81.48	94.44	走向	92.31	95.73	94.02
现在	94.17	95.00	94.17	一定	99.07	98.60	99.07
可以	96.25	95.00	96.25	总会	98.08	98.08	98.08

学会	93.89	89.31	94.66	中长期	92.86	91.07	92.86

表 2 模型测试结果

对于 RFR_SUM，在统计前后从分从合语境词频时，考虑到窗口中的词与歧义词的距离，离歧义词越近的词，赋予较大的词频。我们设定窗口大小为 5，离歧义词按照距离由近到远每次加的词频分别为：5、4、3、2、1。比单纯未考虑距离远近时的效果有所提高。词语相似度计算中的参数参照文献[9]的取值；另外，参数 KnnNum 取 3，测试结果见表 2。本实验中我们使用的是知网 2000 版本，所以很多上下文窗口中的词，该版本知网中并没有收录，也因此导致在计算语境相似度时有误差。三个模型对测试句子逐一都进行了处理，所以精确率、召回率、F 值三者相等，其中 RFR_SUM 模型、语义计算模型的平均 F 值分别为 96.01%、93.42%，将两者结合的模型的平均 F 值达到了 96.52%。

从表中可以看出，RFR_SUM 的总体效果不错。而基于知网的语义相似度计算模型的效果不是很理想，只有“马上”、“现在”、“正当”、“中将”、“走向”的 F 值高于 RFR_SUM 的测试结果。其它 15 个均低于或等于 RFR_SUM 模型结果。可见，我们单纯利用知网从语义相似度方面出发来消解歧义是不够的。因此，我们将两个模型结合起来，效果却比单个 RFR_SUM、语义相似度计算模型都有了提高，实验的最终平均 F 值比 RFR_SUM 高出 0.51%，高于语义相似度计算模型 3.10%。

3.3 两模型结合效果与其它模型效果对比

为了验证以上两个模型的结合性能，我们分别进行了 CRF 模型(即条件随机场模型)[12]和最大熵模型[13]的测试实验。

首先，为验证效果，我们选取与 RFR_SUM 模型实验用的相同特征，即仅以词作为特征建立模板，CRF 测试后的平均 F 值为 94.90%。CRF 模型作为一个较好的通用型序列标注模型，应用于自然语言处理的各个领域，大都取得了理想的效果，但在大致相同的特征下，都不运用词性等句法信息，CRF 平均 F 值低于表 2 中“两者结合”的结果 1.62%。

然后，我们增加相应特征建立模板，使用上下文窗口为 2 的词、词性、词与词共现、词性与词性共现作为特征[14]。最优特征下 CRF 测试结果如表 3：

才能 100.00	都会 97.58	个人 99.52	将来 97.67	马上 100.00
人才 96.94	上来 93.95	现在 95.00	可以 94.34	学会 91.95
一块 92.86	一起 98.14	正当 95.30	中共 99.12	中学 96.19
中将 91.14	走向 97.44	一定 98.60	总会 99.03	中长期 91.07
平均 F 值:	96.30			

表 3 最优特征下 CRF 测试结果

从表中可以反映出增加特征后的平均 F 值比先前提高了 1.4%，但仍低于我们的模型 0.22%。

此外，我们还利用最大熵模型进行了比对实验，选用词、词性、词与词性的共现作为其特征，经过试验将上下文窗口设为 5 时，最大熵平均 F 值最大，其测试结果如表 4：

才能 94.30	都会 99.19	个人 97.43	将来 96.51	马上 100.00
人才 91.30	上来 87.04	现在 95.00	可以 93.75	学会 87.02
一块 96.43	一起 95.45	正当 98.68	中共 99.12	中学 97.22
中将 93.83	走向 91.45	一定 99.53	总会 98.08	中长期 91.07

平均F值:	95.12

表4 最大熵测试结果

最大熵测试结果仍逊于我们的模型，可见我们的模型不仅需要的语言学特征少，需要的语料的深加工程度低，而且效果甚佳。

4 总结与展望

RFR_SUM 模型将语境的相对词频相加的过程，类似于人们根据语境信息进行综合决策的过程。模型较好地模拟了人类分词消歧的过程，具有本真性，因此取得了较好的效果。但该模型对于词语间的语法制约关系考虑的不多，因此我们利用知网对概念的表示，通过计算语境的语义相似度，对学习到的共现词语进行语义归纳，合理地与 RFR_SUM 模型相结合，提高了组合型歧义消歧效果，高于常用的 CRF 和最大熵模型。

我们的下一步工作主要是：(1) 充分利用语言学资源，改进语境的语义相似度计算模型，提高消歧效果。(2) 改进基于知网语义相似度的计算方法。(3) 寻求 RFR_SUM 模型与语境相似度计算模型更有效的结合方法，优劣互补，充分发挥两者的优势，并尝试扩大该模型的应用范围。

参 考 文 献

[1] 刘开瑛, 由丽萍. 汉语框架语义知识库构建工程[A]. 中文信息处理前沿进展[C]. 北京：清华大学出版社, 2006:64-71.

[2] 孙茂松，黄昌宁，邹嘉彦. 利用汉字二元语法关系解决汉语自动分词中的交集型歧义[J]. 计算机研究与发展, 1997, 34(5):332-339.

[3] 孙茂松，左正平. 消解中文三字长交集型分词歧义的算法[J]. 清华大学学报，1999, 39(5):101-103.

[4] 刘扬, 于江生, 俞士汶. CCD构造模型及VACOL辅助软件的设计与实现[J]. 语言文字应用, 2003(1):83-88.

[5] Qu Weiguang, Sui Zhifang, et al. A collocation-based WSD model: RFR-SUM[J]. 2007, LNAI, 4570:23-32.

[6] 曲维光. 现代汉语词语级歧义自动消解研究[M]. 北京:科学出版社, 2008.

[7] 刘开瑛. 中文文本自动分词和标注[M]. 北京:商务印书馆, 2000.

[8] 董振东, 董强. 知网[DB/OL], http://www.keenage.com.

[9] 刘群, 李素建. 基于《知网》的词汇语义相似度的计算[A]. 第三届汉语词汇语义学研讨会[C], 台北, 2002.

[10] 肖云, 孙茂松, 邹嘉彦. 利用上下文信息解决汉语自动分词中的组合型歧义[J]. 计算机工程与应用, 2001, 37(19):87-81.

[11] John Lafferty, Andrew McCallum, et al. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data[A]. Proc. International Conference on Machine Learning[C], 2001.

[12] Adam L.Berger, Stephen A. Della Pietra, et al.A Maximum Entropy Approach to Natural Language Processing [J]. Computational Linguistic, 1996, 22(1).

[13] 丁德鑫, 曲维光, 徐涛等. 基于 CRF 模型的组合型歧义消解研究[J]. 南京师范大学学报(工程技术版), 2008, 8(4):73-76