# 基于语言模型验证的词义消歧语料获取*

郭宇航 车万翔 刘挺
哈尔滨工业大学信息检索研究室, 哈尔滨。 150001
E-mail: {yhguo, car, tliu}@ir.hit.edu.cn

摘要：作为一种稀缺资源，人工标注语料的匮乏限制了有指导词义消歧系统的大规模应用. 有人提出了利用目标词的单义同义词在生语料中自动获取词义消歧语料的方法，然而，在某些上下文当中，用目标词替换这些单义的同义词并不合适，从而带来噪声. 为此，我们使用语言模型过滤这些噪声，达到净化训练数据，提高系统性能的目的. 我们在Senseval-3 中文 lexical sample 词义消歧数据集上进行了实验，结果表明经过语言模型过滤的词义消歧系统性能明显高于未经过滤的系统.

关键词： 词义消歧；互联网语料；单义同义词；噪声；语言模型.

# WSD Corpus Acquisition with Language Models Validation*

Yuhang Guo Wanxiang Che Ting Liu
Harbin Institute of Technology, Harbin 150001, China
E-mail: {yhguo, car, tliu}@ir.hit.edu.cn

**Abstract:** The lack of manually annotated training data is a critical problem faced by supervised word sense disambiguation (WSD) systems. The monosemous lexical relatives substitution of target words have been proposed to acquire WSD corpus from the Web automatically. However, in some cases, the monosemous lexical relatives cannot be substituted by the target word suitably and then noises will be brought in. We propose a language models validation method to filter these noises, which can purify the training data, and improve the performance accordingly. Out experiments on Senseval-3 Chinese lexical sample task show that the system based on the training data acquired from the Web with language models validation achieves better accuracy than the one without language models validation.

**Key words:** word sense disambiguation; Web corpus; monosemous; noise; language model.

## 引言

自然语言中很多词有不只一个含义，词义消歧(Word Sense Disambiguation, WSD)的任务就是要确定这些多义词在特定上下文中的正确词义. 词义消歧是自然语言处理当中的重要问题，对机器翻译[1]，信息检索[2]等领域很有帮助.

基于人工标注语料的有指导词义消歧是当前最为流行词义消歧方法，不过这种方法依赖于

* 本研究得到国家自然科学基金面上项目(60575042，60675034)和 863 项目(2006AA01Z145)资助

* Supported by the National Natural Science Foundation of China under Grant No. 60675034, and No. 60575042; Supported by the National 863 Project under Grant No. 2006AA01Z145

足量的人工标注语料，其获取是比较困难的. 为了在覆盖更多的多义词同时达到较高的系统性能，语料获取这一瓶颈是不得不克服的. 很多人在不同的侧面进行了尝试，包括从互联网上自动获取语料的方法[3]，bootstrapping 方法[4]，利用双语平行文本的方法[5][6]等等. 本文主要关注从互联网自动获取词义消歧语料的方法.

Leacock 等人首先提出以单义同义词作为目标多义词相应词义的方式从语料库中抽取样本作为训练数据的思想[7]，Mihalcea 和 Moldovan 扩展了这个思想，并把把它用在互联网语料库上[8]. Agirre 和 Martínez 在有指导词义消歧上使用了类似方法获得的语料[9]，然而结果却并不理想，他们推断这是由样本数目不均衡所造成的，这一点在他们的后期工作中得到了验证[10].

以上的所有方法都基于这样的假设：单义同义词的上下文必须与其目标词相应词义的上下文相似. 于是只要把单义同义词用目标词替换掉，就可以得到新的训练样本. 但这样的假设并非无可挑剔，尤其是当这种词替换并不合适的时候. 例如，作为"understand"的含义，中文“理解”是目标词“把握”的单义同义词，然而，句子“尤纳斯 希望 大家 要 理解 姚明 。”中的“理解”却不适合替换成为“把握”，也就是说，这个句子不适合作为目标词“把握”的训练样本. 如果有很多类似的句子被加入到训练语料中去，就会带来大量的噪声，从而影响最终的词义消歧效果.

为了解决上述问题，我们提出通过语言模型的验证过程来过滤有噪声的样本. 过程分三个步骤：1）首先检索出上千个包含目标词的句子；2）用这些句子建立一个语言模型；3）计算所获取的样本在这个语言模型上的概率，依此过滤噪声.

在 Senseval-3[1]中文 lexical sample 评测集上进行的实验结果表明，使用了语言模型验证过程的有指导词义消歧系统性能明显好于没有这一过程的系统.

## 1 语言模型验证

在这一部分中，我们首先介绍从互联网获取词义消歧语料的基本方法，而后分析这种方法的不足，并给出我们的解决方案——语言模型验证.

### 1.1 从互联网获取词义消歧语料

我们首先人工收集了 Senseval-3 中文 lexical sample 评测集中的每个目标词的单义同义词，这种方法受到了卢等人的启发[11]. 表 1 列出了一个目标词的同义词及相应词义的训练语样本数.

**表 1 Senseval-3 中文 lexical sample 中的目标词：“材料”的单义同义词及利用这些同义词从互联网上获取的训练样本数量**

| 目标词 | 同义词 | 同义词个数 | 获取的训练样本数 |
|---|---|---|---|
| 材料 | 数据，资料，素材，文件，题材 | 5 | 2,153 |
| | 建筑材料，装修材料，原材料，原料，物资，设备 | 6 | 3,121 |

训练样本的获取则分为以下几个步骤：

1：使用 Google[2]从互联网上检索每个单义同义词得到大约 2,000 个 snippets.

[1] http://www.senseval.org

[2] http://www.google.com

2：从 snippet 中抽取包含单义同义词的句子.

3：将其中的单义同义词替换成相应的目标词.

### 1.2 语言模型验证

在以往类似的研究中，获得的所有扩展语料都直接被当作训练语料. 其中的一些错误的替换必然会带来许多噪声，不管同义词和目标词有多么相似，这都是难以避免的. 因此我们必须从所获得的候选样本中区分出噪声样本并过滤掉它们. 这里提出了一条假设：所有适合替换的样本都是好的训练样本. 这样问题就转化成了如何来衡量一个替换是否是合适.

统计语言模型，也称作 N-gram 模型[12]，可以作为解决词替换问题的工具[13]. 它在自然语言处理的很多方向上都有成功的应用，比如语音识别，机器翻译等等. 一般来说，语言模型的使用分两个阶段：首先在一个相对较大的句子集合上训练获得语言模型所需的参数，而后，根据得到的语言模型来估计一个新的句子出现的概率值.

图 1 展示了应用在词义消歧语料的获取任务上的语言模型验证系统结构. 在基本的语料获取系统的基础上，这里加入了语言模型的验证过程(图中虚线部分表示).

## 2 评测的实验设置

这部分中，我们介绍实验用到的人工标注语料和当前流行的有指导词义消歧方法，包括 SVM 分类器及用来表示知识源的特征. 最后介绍一下用到的语言模型工具.

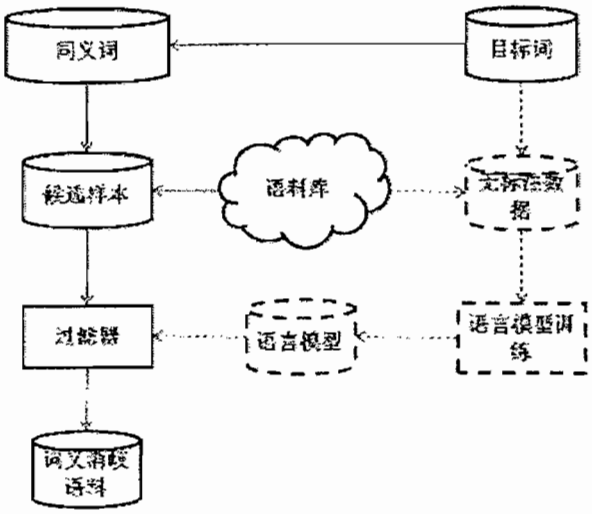


图 1 语言模型验证系统结构

### 2.1 语料

我们使用 Senseval-3 中文 lexical sample 中的语料用于评测. 其中包含了 20 个目标词. 目标词词义的标注依据 HowNet[3][14]知识库中的编码. 对于每个目标词，训练数据提供了 20-100

[3] http://www.keenage.com/

个词义分布均匀的平衡样本. 测试数据的样本数量大致上为训练数据的一半.

### 2.2 有指导词义消歧系统

首先, 我们使用了郭等人的方法建立了一个有指导词义消歧系统[15]. 这一有指导方法被用来评价并对比基于不同训练数据的系统性能, 包括人工标注训练语料, 加入了经过语言模型验证的互联网语料, 以及加入未经过语言模型验证的互联网语料等. 支撑向量机(Support Vector Machine, SVM)[16]以其在词义消歧上的优良性能被选作分类器. 由于SVM是一个二元分类器, 为了解决多类分类问题, 这里采用了一对一的策略将 SVM 推广成为多元分类器. 我们选择Libsvm[17]作为SVM的实现, 使用了默认的C(=1)和e(=0.001)参数, 并选用线性核函数.

令$[...,w_{-3},w_{-2},w_{-1},w,w_{+1},w_{+2},w_{+3},...]$表示目标词$w$的上下文, $p_i$表示$w_i$的词性, $h$为$w$的依存语法父节点, $\{ch_1,...,ch_n\}$为$w$依存语法的子节点之集.如下两种类型的特征被用来表示词义消歧的知识源:

- 局部特征: 词性, 搭配以及句法信息

$$p_{-3},p_{-2},p_{-1},p,p_{+1},p_{+2},p_{+3},w_{-1},w_{+1},(w_{-2},w_{-1}),(w_{-1},w_{+1}),(w_{+1},w_{+2}),h,\{ch_1,...,ch_n\}$$

- 主题特征: 词袋(the bag of words), 句子中出现的 m 个词的上下文之集

$$\{c_1,...,c_m\}$$

### 2.3 语言模型的实现

实验使用 IRST 语言模型工具箱[18]作为语言模型的实现. 在我们的实验中, 将其设置为3-gram 和 Witten-Bell[19]平滑策略.

首先利用从互联网上获得的无标注数据对每个目标词训练语言模型. 再用语言模型计算每个从互联网上由单义同义词获取的句子的混乱度. 混乱度和概率值的关系如下:

$$PP = P(w_1w_2...w_N)^{-\frac{1}{N}}$$

其中$P(w_1w_2...w_N)$是长为$N$的测试字串$w_1w_2...w_N$的概率, $PP$是其混乱度.

在实验中我们用IRSTLM所输出的PPwp作为句子的混乱度. 不同于通常的混乱度, PPwp考虑到了未登录词(Out-Of-Vocabulary, OOV).

## 3 评测与讨论

在这部分中, 我们对比三个基于不同词义消歧语料的系统性能: 人工标注语料, 从互联网上获取的未经过和经过语言模型验证的语料. 在所有的实验中, 我们用准确率作为为评价系统的指标. 微观平均值(micro-average)用来指示系统最终的性能.

### 3.1 人工标注语料

表2的第2列列出了使用Senseval-3中文lexical sample中的人工标注训练语料作为训练数据, 使用有指导词义消歧方法对其中的测试数据进行评测得到的结果, 由于受篇幅所限, 只列出一部分结果.

请注意我们并没有特别地对系统调试参数. 这些参数都是来自于文献中提到的设置以及SVM实现中的默认设置. 同样的设置方式被使用在我们的其他实验当中. 系统的平均准确率达

到了 62.00%. 这和 Senseval-3 当时评测的较好成绩(60.40%)[20]是相当的.

表 2 测试数据集上的系统准确率(%)对比：人工标注数据，全部互联网语料，带有 Senseval-3 偏置的互联网语料以及经过语言模型验证带有 Senseval-3 偏置的互联网语料

| 目标词 | 人工标注 | 全部互联网语料 | 带有 Senseva-3 偏置的互联网语料 | 带有 Senseval-3 偏置且经过语言模型验证的互联网语料 |
|---|---|---|---|---|
| 包 | 52.78 | 55.56 | 52.78 | 61.11 |
| 把握 | 53.33 | 46.67 | 46.67 | 46.67 |
| 材料 | 60.00 | 80.00 | 90.00 | 90.00 |
| 冲击 | 61.54 | 61.54 | 53.85 | 46.15 |
| 平均值 | 62.00 | 56.73 | 57.52 | 58.31 |

## 3.2 互联网语料库

从互联网上由单义同义词获取的全部训练样本被用于训练另一个有指导词义消歧系统. 在测试数据上的准确率如表 2 的第 3 列所示. 可以看到其准确率要低于人工标注的那组. 受到 Agirre 和 Martínez 的启发[10], 我们也考虑了训练数据的偏置问题. 因为已知 Senseval-3 中文 lexical sample 任务中训练和测试数据的词义分布情况: 目标词的每个词义对应的训练样本数都是平衡的, 所以我们也可以构建有相同分布的互联网语料库.对于目标词的每个词义, 我们都保留相同数量的样本数目, 定为所有词义中得到最少语料的词义所获得的样本数. 对于那些多于这一数目的词义, 我们随机地选择超过的部分予以抛弃以保证最终词义的分布和 Senseval-3 中的一致. 系统的最终结果如表 2 中的第 4 列所示.

## 3.3 语言模型验证

首先, 基于由目标词从互联网上获取的句子, 我们对每个目标词训练语言模型, 而后, 对于由单义同义词从互联网上获取的句子, 利用语言模型所提供的评价方式过滤掉低质量部分.

此处依然保持了 Senseval-3 的偏置. 即保证目标词的每个词义对应的训练数据的量是平衡的. 和上一节中不同的是, 我们根据训练样本的混乱度对其排序, 保留具有较小混乱度 的样本. 表2的最后一列给出了此时的系统准确率. 经过语言模型验证的系统性能不仅高于使用了全部互联网数据的系统, 也要好于带有 Senseval-3 偏置的系统. 不过其性能还是要低于使用人工标注语料的结果.

为了测试语言模型方法的潜力, 我们引入两条准则来调节每个目标词的训练样本数目. 这里采用 Senseval-3 中文 lexical sample 训练数据作为开发集. 两条准则分别是: C1: 语言模型所产生的混乱度; C2: 目标词每个词义对应训练样本数的最大值.

根据 C1, 我们设置了一个阈值 $T_{pp}$ .如果一个样本在语言模型上产生的混乱度高于这个阈值, 那么这个样本就会被过滤掉. 图 2 给出了 4 个目标词的准确率受 $T_{pp}$ 的变化的影响.

如图 3 所示, C2 带来系统性能的变化和 C1 带来的变化比较相似. 然而, 由于 C2 避免了偏置问题, 能够获得更高的准确率. 最终系统的优化准确率为 64.64%, 要高于基于 C1 的系统性能(63.06%)甚至显著地 (paired t-test, p=0.05)高于基于人工标注语料的系统(62.00%).

## 4 结论与未来工作

本文提出了用语言模型验证的方式过滤基于单义同义词从互联网上获取的低质量候选训练样本的方法. 在 Senseval-3 中文 lexical sample 数据集上的实验表明该方法能够净化训练数据, 提高系统性能, 甚至能明显地好于单纯使用人工标注语料的系统. 我们可以很容易地对这种方法进行扩展, 并用到全文词义消歧的任务中去.

在今后的工作中, 为了将语言模型验证的方法应用到全文词义消歧, 我们将使用类似于 WordNet 的同义词词典, 自动地获取单义同义词.

此外, 除了语言模型, 还可以尝试其他验证从互联网上获取文本的质量的词替换方法[21], 比如浅层语义分析(Latent Semantic Analysis, LSA), Web mining 等.

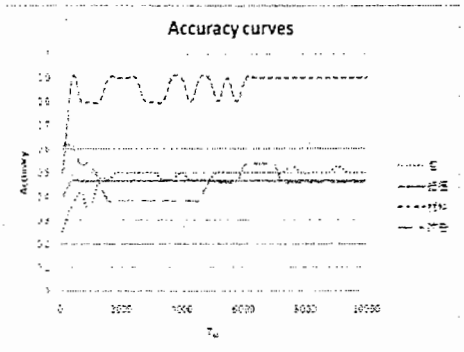


图 2 四个目标词随 $T_{pp}$ 变化的系统性能曲线

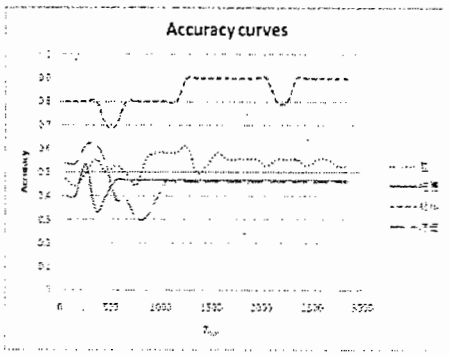


图 3 四个目标词随 $T_{num}$ 变化的系统性能曲线

## 参 考 文 献


[1] Yee Seng Chan, Hwee Tou Ng, and David Chiang. 2007. Word sense disambiguation improves statistical machine translation. In Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, pages 33–40, Prague, Czech Republic, June.

[2] Liqi Gao, Yu Zhang, Ting Liu, and Guiping Liu. 2006. Word sense language model for information retrieval. In AIRS, pages 158–171.

[3] Rada Mihalcea and Dan I. Moldovan. 1999. An automatic method for generating sense tagged corpora. In AAAI '99/IAAI '99: Proceedings of the sixteenth national conference on Artificial intelligence and the eleventh Innovative applications of artificial intelligence conference innovative applications of artificial intelligence, pages 461–466, Menlo Park, CA, USA.

[4] David Yarowsky. 1995. Unsupervised word sense disambiguation rivaling supervised methods. In Proceedings of the 33rd annual meeting on Association for Computational Linguistics, pages 189–196, Morristown, NJ, USA.

[5] Yee Seng Chan and Hwee Tou Ng. 2005. Scaling up word sense disambiguation via parallel texts. In Manuela M. Veloso and Subbarao Kambhampati, editors, AAAI, pages 1037–1042. AAAI Press / The MIT Press.

[6] Hang Li and Cong Li. 2004. Word translation disambiguation using bilingual bootstrapping. Computational Linguistics, 30(1):1~22.

[7] Claudia Leacock, George A. Miller, and Martin Chodorow. 1998. Using corpus statistics and wordnet relations for sense identification. Computational Linguistics, 24(1):147~165.

[8] Rada Mihalcea and Dan I. Moldovan. 1999. An automatic method for generating sense tagged corpora. In AAAI '99/IAAI '99: Proceedings of the sixteenth national conference on Artificial intelligence and the eleventh Innovative applications of artificial intelligence conference innovative applications of artificial intelligence, pages 461–466, Menlo Park, CA, USA.

[9] Eneko Agirre and David Martínez. 2000. Exploring automatic word sense disambiguation with decision lists and the web. In Proceedings of the Semantic Annotation And Intelligent Annotation workshop organized by COLING Luxembourg 2000.

[10] Eneko Agirre and David Martínez. 2004. Unsupervised wsd based on automatically retrieved examples: The importance of bias. In Dekang Lin and Dekai Wu, editors, Proceedings of EMNLP 2004, pages 25~32, Barcelona, Spain, July.

[11] Zhimao Lu, Haifeng Wang, Jianmin Yao, Ting Liu, and Sheng Li. 2006. An equivalent pseudoword solution to chinese word sense disambiguation. In Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, pages 457~464, Sydney, Australia, July.

[12] C.D. Manning and H. Schütze. 1999. Foundations of Statistical Natural Language Processing. MIT press, Cambridge, MA.

[13] Shiqi Zhao, Ting Liu, Xincheng Yuan, Sheng Li, and Yu Zhang. 2007. Automatic acquisition of contextspecific lexical paraphrases. In IJCAI, pages 1789~1794.

[14] Zhendong Dong and Qiang Dong. 2006. Hownet And the Computation of Meaning. World Scientific Publishing Co., Inc., River Edge, NJ, USA.

[15] Yuhang Guo, Wanxiang Che, Yuxuan Hu, Wei Zhang, and Ting Liu. 2007. Hit-ir-wsd: A wsd system for english lexical sample task. In Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007), pages 165~168, Prague, Czech Republic, June.

[16] Vladimir N. Vapnik. 1998. Statistical Learning Theory. Wiley.

[17] Chih-Chung Chang and Chih-Jen Lin, 2001. LIBSVM: a library for support vector machines.

[18] Marcello Federico and Mauro Cettolo. 2007. Efficient handling of n-gram language models for statistical machine translation. In Proceedings of the Second Workshop on Statistical Machine Translation, pages 88~95. ACL, June.

[19] Ian H. Witten and Timothy C. Bell. 1991. The zerofrequency problem: Estimating the probabilities of novel events in adaptive text compression. IEEE Transactions on Information Theory, 37(4):1085~1094.

[20] Zheng-Yu Niu, Dong-Hong Ji, and Chew-Lim Tan. 2004. Optimizing feature set for chinese word sense disambiguation. In Rada Mihalcea and Phil Edmonds, editors, Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text, pages 191~194, Barcelona, Spain, July.

[21] Samer Hassan, Andras Csomai, Carmen Banea, Ravi Sinha, and Rada Mihalcea. 2007. Unt: Subfinder: Combining knowledge sources for automatic lexical substitution. In Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007), pages 410~413, Prague, Czech Republic, June.