一种基于 N-Best 结果组合优选的词语对齐方法

朱丹青，常宝宝

（北京大学信息科学技术学院，计算语言学研究所，北京 100871）

E-mail: {zhudanqing, chbb}@pku.edu.cn

摘要：在这篇文章中，我们提出了一种从句对齐语料中抽取出词语对齐的新颖方法。我们比较主流的词语对齐方法，重点分析 IBM 模型，发现模型在挑选最佳对齐方面的缺陷。我们对每组对齐取 NBest 的结果，然后利用有监督的学习方法对 NBest 结果进行组合优选，取得不错的结果。

关键词：词对齐；组合优选；有监督学习

A Method of Word Alignment Based on N-Best Results Combined Optimum Choice

Danqing Zhu, Baobao Chang

(Institute of Computational Linguistics, School of Electronic Engineering and Computer Science, Peking University, Beijing, Zip Code: 100871, China)

E-mail: { zhudanqing, chbb }@pku.edu.cn

Abstract: The research of word alignment is brought up with Machine Translation. In this paper, we proposed a new algorithm to improve word alignment. Currently, there were several aspects to improve word alignment, modifying the generative model, using the discriminative model, and adopting* semi-supervised methods. After analyzing the IBM Models, we found the defect of model, when choosing the Viterbi alignment. We got the N-Best results for each alignment, and then made choice by supervised learning, gaining good results.

Keywords: Word Alignment; Combined Optimum Choice; Supervised Learning

1 引言

在自然语言处理这个领域中，双语对齐的研究工作伴随着机器翻译研究深入而不断发展。对齐的结果对机器翻译有着直接的影响。此外，词语对齐在语义消岐、双语词典的建设等方面都起着重要作用。

根据对齐的层次，双语对齐可以细分为：句子对齐，短语对齐和词语对齐。

到了 20 世纪九十年代，词语对齐进入了一个高速发展的时期。随着统计方法在自然语言处理的各个领域中取得瞩目的成果，对齐也在统计方法的指导下有了长足的进步，其标志是上个世纪九十年代初 Brown 等人提出的五个词对齐模型[1]。接着九十年代中期 Vogel 等人在 Brown 的基础上提出了 HMM 词对齐模型[2]。二十一世纪初 Och 等人又在前人的基础上提出了一个新的模型和一种对称化（symmetrization）的方法[3]。这些模型的特点都是生成模型，对语言没有限制，有着很强的适应性。

目前双语词对齐的研究主要集中在英汉和英法之间。研究表明，英法的双语词对齐的效果要远远好于英汉的词对齐。英法的词对齐错误率(Alignment Error Rate, AER)一般能达到 20%以下，而英汉的词对齐错误率（Alignment Error Rate, AER）都维持在 30%左右。

基金资助：国家社会科学基金项目(06BYY048)、国家自然科学基金项目(60303003)

1.1 词对齐研究的现状

进入了二十一世纪，词对齐的发展进入了新的阶段。统计机器学习方兴未艾，其方法大量的应用于各个领域。在这种背景下，词对齐的研究工作主要从三个方面着手：(1)改善原有的生成模型；(2)利用有监督的机器学习方法，构造判别模型；(3)利用半监督的机器学习理念，处理词对齐。

Yonggang Deng[4]和 Yanjun Ma[5]分别在 2007 年提出各自对生成模型的改进方法。Deng 利用一些外部的语言资源对源词和目标词先进行聚类，然后在对齐中加入了类之间的对应关系。Ma 是将对齐关系成连续性的词进行合并，利用 bootstrapping 的方法进行对齐，巧妙地避开了原模型对多词对齐处理的缺陷。

刘洋[6]、Phil Blunsom[7]、Necip Fazil Ayan[8]、Robert C. Moore[9]等人分别在 2005 年到 2006 年提出利用各种判别模型处理词语对齐问题。刘洋和 Ayan 利用最大熵模型去挑选可靠的词对齐。Blunsom 利用条件随机场模型进行对齐方面的研究。Moore 构造一个判别框架，在他的判别框架中，Moore 进行了一系列实验。判别模型优于生成模型在于其能充分利用资源，设计各种特征模板，灵活地利用多种特征。

Alexander Fraser[10]和王海峰[11]分别在 2006 年提出利用半监督的方法去解决词对齐的问题。半监督方法的提出是基于这样事实，已标注的词对齐语料过少导致有监督的方法效果很差，而未标注的词对齐语料很多。这种情况很适合于半监督方法。Fraser 是基于半监督的 EM 算法。王海峰是采用半监督的 Adaboost 算法。

1.2 文章的组织

本文提出了一种从句对齐语料中抽取出词语对齐的新颖方法。我们分析 IBM 对齐模型在选取最佳对齐方面的缺陷，然后利用有监督的学习方法对 NBest 结果进行组合优选。文章接下来的部分是这样组织的：第二部分简要介绍 IBM 对齐模型和存在的缺陷。第三部分详细阐述 NBest 的方法。第四部分说明实验、实验结果，并对结果进行分析评价。最后，我们对所作的工作进行总结，并指出未来研究工作的方向。

2 IBM 对齐模型和 Viterbi 对齐的缺陷

2.1 IBM 模型

IBM 的 Brown 等人提出两种思路，类 HMM 对齐模型和基于繁殖率的对齐模型。[1]

类 HMM 对齐模型的思路是：给定一个源语言的句子，先确定目标语言句子的长度，然后确定与源语言句子的第 i 位置成对齐关系的词在目标语言句子中的位置，最后确定目标语言句子中的词。基于繁殖率的对齐模型的思路是：给定一个源语言词后，考虑这个源语言词一般是由若干个目标语言词翻译过来的，这样的目标语言词数目就是这个源语言词的繁殖度 ϕ。然后，确定由 ϕ 个目标语言词组成的词集。接下来，对词集确定目标语言词在句子中的位置集。

IBM 的模型 1 和模型 2 是基于类 HMM 对齐模型框架，IBM 的模型 3，模型 4，模型 5 是基于繁殖率的对齐模型框架。由于模型框架过于复杂，模型 3 做了最大的简化，模型 4 在模型 3 的基础上进行了一定的泛化，模型 5 又在模型 4 的基础上进行了进一步的泛化。

---

[1]参照 Och[3]的说法

2.2 Viterbi 对齐的缺陷

Viterbi 对齐指的是根据模型计算出来概率最大的对齐 。这里就存在一个问题，我们能否计算出模型中概率最大的对齐。经过论证，模型一和模型二是可以得到模型概率最大的对齐，但模型三之后，由于模型本身设计的缺陷，那个真正的概率最大的对齐是无法得到的，因此，一般采用有缺陷的计算方法得到概率最大的对齐去代替 Viterbi 对齐。IBM 模型由于本身的复杂性，这些模型在计算参数时，就采用了各种简化的方法，而且高一级模型又是以低一级模型为基础的。这样就存在着一个严重错误传递的问题。所以 IBM 模型得到的 Viterbi 对齐与最佳对齐是存在差距的。

3 基于 N-Best 结果组合优选的对齐方法

3.1 N-Best 对齐的效果

根据上面的分析，我们可以试着放大我们取值的范围，除了 Viterbi 对齐外，我们也要考虑概率值仅次于 Viterbi 对齐的 N-1 个对齐，即 N-Best 结果。这样，在 N-Best 的结果中，是不是 Viterbi 的结果就是最好的呢？不妨，我们先作一个大胆的假设，Viterbi 的结果不一定是最好的结果。为了证实这个假设，我们做了如下实验，对 IBM 模型结果取 N-Best，然后计算 N-Best 中的最佳结果。实验结果如下图所示，

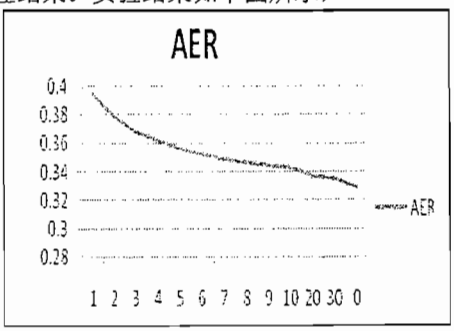

图 1：N-Best 最佳结果的平均错误率

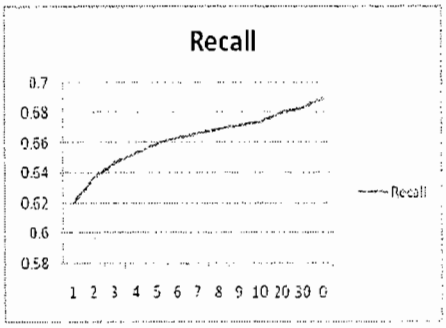

图 2：N-Best 最佳结果的召回率

上两个图横轴坐标为 N-Best 中 N 的取值，N=1 表示为模型输出的 Viterbi 对齐，N=0 表示为取模型的所有对齐结果。可以很明显的看出，随着 N 的增加，对齐的效果在变好，而且趋势是十分明显的，所有对齐结果中最好的 AER 比 Viterbi 对齐结果少了十多个百分点， 召回率也相应高了近七个百分点。因此，我们可以充分地去利用 IBM 模型的 N-Best 对齐结果。

3.2 有监督的判别模型 N 的选择

由于 N-Best 结果中的最佳结果比 Viterbi 结果好，这证明，在一些句对中，Viterbi 结果并不是最佳结果。我们将 N-Best 的结果取并集，从图 2 就可以看出随着 N 的增大，最佳结果的 recall 值上升，则并集中的正确对齐数必然比 Viterbi 结果中多。基于这样的事实，我们力图去找到一种方法，对并集进行筛选，得到一个尽可能好的结果。我采用的方法如下：

1. 将 N 个对齐结果取并集，得到 N-Best 结果；
2. 利用有监督的判别模型中上一步的结果中挑出正确的结果。

a) 构造训练语料，利用人工给出的标准答案，对 N-Best 结果中的每个对齐进行标注，属于标准答案的对齐标记为 Y，不属于标准答案的对齐标记为 N

b) 利用已有的分词信息和词性信息，抽取出特征文件。

c) 利用分类器进行训练，构造出一个判别模型

d) 利用这个判别模型对新语料进行分类，保留属于 Y 类的对齐结果，这个结果就是最终的结果

3.2.1 N 的选择

由于将 N-Best 的对齐结果取并集，如果 N 取值过大，合并的结果中有太多错误对齐，这对最后的结果是有不利的影响的，极端的例子是，N 充分大，合并的结果就可能是一个 的笛卡尔积。经过一系列对折交叉校验实验（每组实验都是 20 次），实验结果见图 6，最后确定当 N 取 4 时，效果最好。

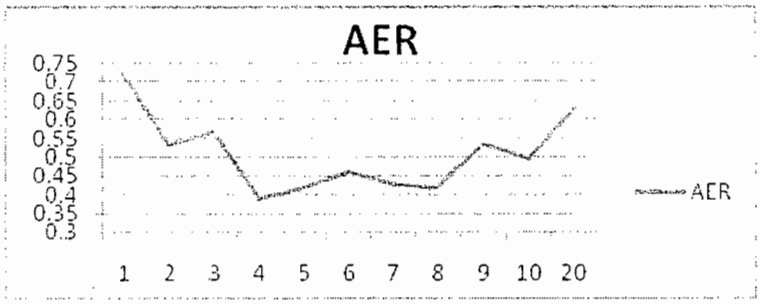

图 6：不同 N 取值的实验结果，横坐标表示 N 的取值，纵坐标表示 AER 的取值

3.2.2 特征模板

每个对齐都涉及到一个源语言词和目标语言词，所以在特征选择时，需要同时涉及到两种语言。我设计的特征模板有：

表 1 备选的特征模板

ID	特征模板	ID	特征模板
1	e_i 的 id+ f_j 的 id	2	e_i 词性+ f_j 词性
3	e_{i-1} 的 id+ f_{j-1} 的 id	4	e_{i-1} 词性+ f_{j-1} 词性
5	e_{i+1} 的 id+ f_{j+1} 的 id	6	e_{i+1} 词性+ f_{j+1} 词性
7	e_{i-2} 的 id+ f_{j-2} 的 id	8	e_{i-2} 词性+ f_{j-2} 词性
9	e_{i+2} 的 id+ f_{j+2} 的 id	10	e_{i+2} 词性+ f_{j+2} 词性
11	e_{i-1}、 e_{i+1} 的 id+ f_{j-1}、 f_{j+1} 的 id	12	e_{i-1}、 e_{i+1} 词性+ f_{j-1}、 f_{j+1} 词性
13	$i-j$[2]		

经过一系列实验，我最后选择的特征模板是 1，2，4， 6，13。这里我选用的特征模板不多。因为实验的数据有限，总共只有 1k 左右的句子，导致大部分特征过于稀疏，对问题的处理产生负面影响。

[2]源语言词和目标语言词在各自句子中位置的跨度

4 实验以及评价

4.1 数据

我们对英汉词对齐语料进行实验。实验采用的数据有两份，一份句对齐语料作为 IBM 模型的训练语料，这部分语料选自北京大学双语句对齐语料库[12]，包括 160，000 个句对，大约有 69，000 个不同英文单词和 62，000 个不同中文单词，英文单词的总数约为 2.89M，中文单词的总数约为 2.78M。另一份是已经有人工标注过词对齐的句对齐语料作为第二阶段实验的训练和测试语料，这部份语料是 2005 年 863 机器翻译评测语料[3]。该语料包括 1007 个句对，大约有 4，300 个不同的英文单词和 4，000 个不同的中文单词，英文单词的总数约为 19.5K，中文单词的总数约为 18.5K。

4.2 工具

在实验中，我们尽可能的在简单可靠的基础上利用到语言上的各种知识。考虑到现阶段自然语言处理的状况，我们决定利用语言中的分词信息和词性信息。

对于英文，分词方面，我们参考了一些资料，然后实现了一个分词程序。词性标注方面，我们采用东京大学的 postagger-1.0[4]进行词性标注。

对于汉语，我们采用北京大学计算语言学研究所的汉语词语切分与词性标注软件[5]进行分词和词性标注。

4.3 评价指标

众所周知，人工进行词对齐是一个复杂而又存在歧义的工作。故我们应用了一个标记系统来接受这种歧义。标记将对齐分成两类：（sure）对齐，没有歧义的对齐；（possible）对齐，可能存在歧义的对齐；。

一个对齐的质量由近似重新定义的准确率和召回率来表示：

$$precision = \frac{|A \cap P|}{|A|} \tag{1}$$

$$recall = \frac{|A \cap S|}{|S|} \tag{2}$$

由著名的 F 值推导出来的对齐错误率（AER）：

$$AER(S, P; A) = 1 - \frac{|A \cap S| + |A \cap P|}{|A| + |S|} \tag{3}$$

4.4 实验

我们的实验分为两大阶段。第一步是利用 160，000 个句对齐语料训练出 IBM 模型 4，然后

---

[3]这部分语料由中科院计算所刘群博士提供，特此致谢。

[4] http://www-tsujii.is.s.u-tokyo.ac.jp/

[5] http://icl.pku.edu.cn/icl%5Fres/segtag98/

对 1007 句语料进行双向（中文到英文和英文到中文）词对齐，得到一组 Viterbi 对齐结果和一组 N-Best（N 个对齐结果取并集）对齐结果。然后我们利用“refined”的方法将每一组中两个方向的对齐结果进行合并。（“refined”的方法具体请参考 Och 于 2000 年发表的文章）。这样得到两个结果。一个是 Viterbi 对齐结果，我们用这个结果作为 baseline。另一个是 N-Best 的结果，我们下一个阶段的实验就是在这个结果上进行的。用来训练这个 baseline 的工具就是著名的 GIZA++工具包[6]。

第一阶段中我们取 N 为 4，这是由一系列实验得出的经验数据。第二阶段，实验的思路就是构造出一个判别模型，从合并的 N-Best 结果找出那些正确的对齐。这个阶段的语料都是已标注的，采用严格的对折交叉检验来证实这个方法。

在我们的实验中，我们采用的是最大熵的方法，工具是 zhang le 开发的最大熵分类器 。该分类器实现了包含高斯平滑在内的最大熵算法，可以很方便地处理分类问题。

4.5 实验结果

表 2 实验结果

	Precision Value	Recall Value	AER
Baseline	0.714432	0.618992	0.316076
Our Method	0.809827	0.481447	0.390083

从结果中，我们可以看出，我们的新方法虽然最后结果在 AER 上比 baseline 差了不少，关键在召回率上差了太多，近 13 个百分点，而在准确率上，这种方法有了显著的提升，从 71%上升到 80%。所以该方法在挑出正确的对齐方面有很大的帮助。

我们在第二阶段进行了 20 次对折交叉检验。从下面的 20 次对折交叉检验的结果不难看出。这种在准确率上面的提升是很稳定的。

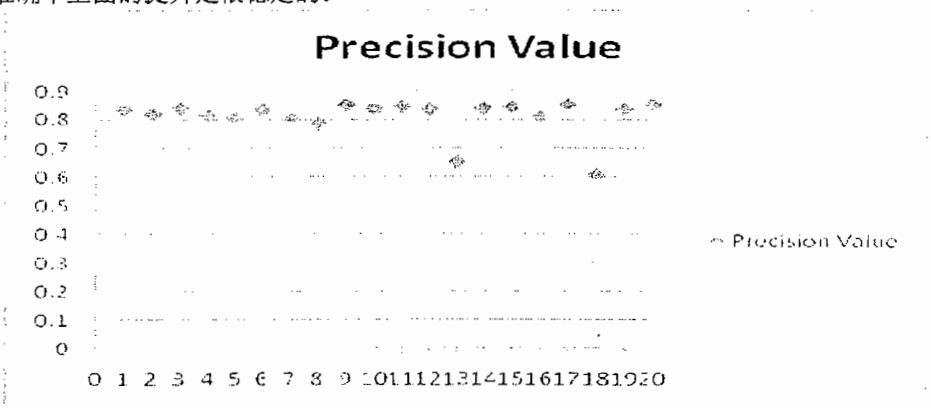

图 7：20 次对折交叉检验结果的准确率

这种方法在召回率上的糟糕表现在很大程度上是由于训练语料过少，导致数据过于稀疏。这也从一个侧面看出有监督的机器学习方法的局限性，即当训练语料很少时，有监督的判别模型不如无监督的生成模型。

[6] http://www.fjoch.com/GIZA++.html

5 结论和将来的工作

我们这次的研究就是尝试着利用 IBM 模型的结果，将对齐问题转化为分类问题。利用比较成熟的分类解决方法，去努力提高对齐的结果。虽然最后的结果不理想，但在准确率这个方面，我们做到了提高。

下一阶段的研究，我们还是继续着眼于将对齐问题进行形式上的转化，充分地去利用现阶段比较成熟的机器学习方法去解决词对齐问题。

参 考 文 献

[1] P. F. Brown, S. A. Della Pietra, V. J. Della Pietra, and R. L. Mercer. 1993. The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics*, 19(2): 263-311

[2] S. Vogel, H. Ney, and C. Tillmann, 1996. HMM based word alignment in statistical translation. In COLING, pages 836-841, Copenhagen, Denmark

[3] F. Och and H. Ney, 2003. A systematic comparison of various statistical alignment models. Computational Linguistics, 29(1):19-51

[4] Y. Deng and Y. Gao. 2007. Guiding Statistical Word Alignment Models With Prior Knowledge. In Proceedings of ACL, pages 1-8, Prague, Czech Republic

[5] Y. Ma, N. Stroppa, and A. Way. 2007. Bootstrapping Word Alignment via Word Packing. In Proceedings of ACL, pages 304-311, Prague, Czech Republic

[6] Y. Liu, Q. Liu, and S. Lin. 2005. Log-linear models for word alignment. In Proceedings of ACL, pages 459-466, Ann, Arbor

[7] P. Blunsom and T. Cohn. 2006. Discriminative word alignment with conditional random fields. In Proceedings of COLING/ACL, pages 65-72

[8] N. Ayan and B. Dorr. 2006. A maximum entropy approach to combining word alignments. In HLT-NAACL, New York, USA

[9] R. C. Moore. 2005. A discriminative framework for bilingual word alignment. In Proceedings of HLT-EMNLP, pages 81-88, Vancouver, Canada

[10] A. Fraser and D. Marcu. 2006. Semi-supervised training for statistical word alignment. In Proceedins of COLING/ACL, pages 769-776

[11] H. Wu and H. Wang 2006. Boosting statistical Word Alignment Using Labeled and Unlabelede Data. In Proceedings of COLING/ACL, pages 913-920

[12] 常宝宝、柏晓静，北京大学汉英双语语料库标记规范，《汉语语言与计算学报》，2003 年 6 月，第 13 卷 2 期，195-214