多策略的机器翻译

王海峰

东芝（中国）研究开发中心　北京　100738

wanghaifeng@rdc.toshiba.com.cn

摘要：统计机器翻译方法由于其良好的数学模型、无指导的学习能力、良好的鲁棒性等优点而备受重视。基于规则的机器翻译方法对语言规律良好的概括和描述能力以及基于实例的机器翻译方法对相似句子相当精确的翻译也颇有价值。多种方法相融合的机器翻译策略正在成为机器翻译发展的重要方向。本文全面介绍东芝对规则、实例、统计等各种机器翻译方法的研究，以及在融合各种方法方面的探索。

关键词：机器翻译，基于规则的机器翻译，基于实例的机器翻译，统计机器翻译，多策略机器翻译

Multi-Strategy Machine Translation

Wang Haifeng

Toshiba (China) Research & Development Center　Beijing　100738

wanghaifeng@rdc.toshiba.com.cn

Abstract: Statistical Machine Translation (SMT) is very popular for its adequate mathematical model, unsupervised learning capacity, and robustness. Rule-Based Machine Translation (RBMT) and Example-Based Machine Translation (EBMT) methods are also valuable. Rules are good at modeling linguistic theory and phenomena. EBMT method is able to translate similar input very well. Multi-strategy machine translation becomes more and more important in the machine translation community. This paper introduces Toshiba's efforts on RBMT, EBMT, SMT and the combination of the three methods.

Keywords: Machine Translation, Rule-based Machine Translation, Example-based Machine Translation, Statistical Machine Translation, Multi-Strategy Machine Translation

1 引言

在三种主流机器翻译方法中，统计机器翻译方法正受到学术界更多的关注。一方面，在具有广泛影响力的 NIST 机器翻译评测中[1]，统计机器翻译系统连连胜出；另一方面，在 ACL 等主流学术会议上，也是多数机器翻译方面的论文都集中在统计方法上。那么，是不是基于规则的机器翻译方法和基于实例的机器翻译方法就已经江河日下，行将退出历史舞台了呢？让我们首先关注一下规则方法和实例方法的现状。

规则方法虽然已经较少出现在主流学术会议上了，但一个有趣的现象是，目前真正为用户所使用的机器翻译产品大多都是基于规则的。同时，规则方法也并不是停滞不前了，而是仍有不断提高的潜力[1]。

目前来看，实例方法不如规则方法在产品中应用广泛，也不如统计方法在学术会议上广受重视及在评测中胜出，但实例方法也依然活跃并富有潜力。在 2005 年的机器翻译峰会（MT Summit X）上，实例方法仍然是与会学者大量讨论的话题，而且 MT Summit X 还专门组织

[1] http://www.nist.gov/speech/tests/mt/

了一个基于实例机器翻译的 Workshop，著名的机器翻译专业国际期刊“Machine Translation”也将在 2006 年下半年发行基于实例机器翻译的专刊。从译文质量的角度，实例方法也毫不逊色甚至还能做得更好[2, 3, 4]。

虽然统计方法由于其良好的数学模型、无指导的学习能力、良好的鲁棒性等优点而备受重视，但规则方法对语言规律良好的概括和描述能力及实例方法对相似句子相当精确的翻译也无疑是颇有价值的。事实上，多种方法相融合的机器翻译策略正在成为机器翻译发展的重要方向。例如，基于语法的模型越来越多的出现在统计机器翻译的研究中[5, 6, 7, 8]；Groves 和 Way 提出了 Statistical EBMT[9] 和 Example-based SMT[3]的概念，等等。而更多的机器翻译系统则是在一种翻译策略为主体的前提下，在系统的不同模块分别使用了不同的技术，例如[4]就在基于实例的机器翻译框架下使用了基于规则的分析和基于统计的生成。

本文将全面介绍东芝在机器翻译领域的研究，包括基于规则的、基于实例的、统计的机器翻译方法，以及各种方法的融合。

2 东芝的机器翻译研究

2.1 概述

从最初开展基于规则的机器翻译研究，到现在全面研究规则、统计、实例等各种方法，东芝在机器翻译方面的研究开发已有二十多年的历史，并基于多年的研究成果开发出了一系列的机器翻译产品。其产品形态包括：机器翻译软件包、机器翻译引擎授权、翻译服务、硬件产品预装等。翻译的语言包括中、日、英三种语言六个方向的互译。

2.2 基于规则的机器翻译

2.2.1 基于转换的翻译方法

东芝的基于规则的机器翻译系统中，采用的是一种基于转换的方法[10]。其主要翻译过程包括分析、转换，生成三部分。其中，分析过程包括词法分析、语法分析和语义分析等三个子过程；转换过程包括预转换、词汇转换和结构转换等三部分；生成则包括词法生成和句法生成。

如图 1 所示，一个待翻译的源语言句子首先被分析为表示其语义结构的依存树，然后源语言的树被转换为目标语依存树，最后生成模块会根据目标语依存树生成目标语译文。

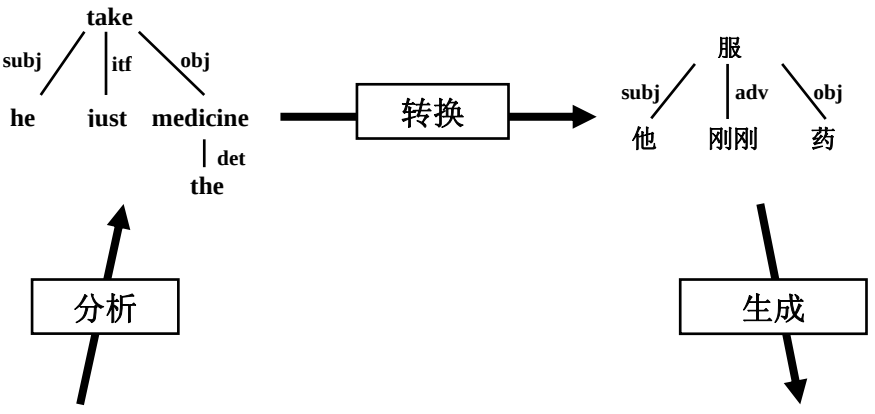

图 1　基于转换的翻译过程

2.2.2 规则系统的组织

使用基于规则的机器翻译方法，很容易开发出一个小型实验系统，但在开发面向真实应用的大型系统时，问题则要复杂得多。在基于规则的机器翻译系统中，随着系统规模不断扩大，规则之间相互影响甚至冲突的可能性也会随之增加，而这种相互影响和冲突将使系统性能的提升变得越来越困难，甚至是步履维艰、停滞不前。

在东芝的基于规则的机器翻译系统中，规则的组织是多层次、细粒度的。如2.2.1节所述，系统主要包括分析、转换、生成三个模块，每一部分又包括一些子模块。更细的，一个子模块还会被分为几个不同的层次。系统中，规则是被分层次调用的，不同层次之间的规则不会被交叉调用。在规则层次被细化的同时，规则本身的知识粒度也被细化。系统中，多数规则都是描述一个具体的、精细的语言现象的，而且是附着在具体词上的。通过这种多层次、细粒度的组织，规则之间的相互影响与冲突得到了有效控制，实现了良好的可扩展性。

2.2.3 可扩展性

为了研究系统的可扩展性，我们定量分析了规则数量与翻译系统性能之间的关系[1]。分析结果表明，翻译系统的性能几乎是随着规则数量的增加而线性增长的。于是，我们得到了一个线性模型来描述规则数量与翻译系统性能之间的关系，其统计相关系数在 0.98 以上。进一步实验表明，这个线性模型对规则数量与翻译系统性能之间的关系不但有非常好的描述能力，而且有很好的预测能力，随着规则的继续增加，翻译系统性能将沿着这条曲线继续增长。这意味着，我们的基于规则的翻译系统拥有良好的可扩展性，仍有不断提高的潜力。

2.3 基于实例的机器翻译

近年来，越来越多的基于实例的机器翻译系统将翻译实例表示为带标注的树结构[11, 12, 13]。在这类系统中，翻译实例中源语言和目标语言句子的分析树，以及两种语言的子树的结构对应关系都存储在实例库中。引入这种带标注的树无疑会提高翻译系统的性能，但这种树-树对应的方法也有其问题。一方面，一些语言缺少高质量的分析工具；另一方面，对于语言结构差异较大的语言对，结构对应关系是很难建立的[14]。因此，我们采用了树串映射来存储翻译实例，并在此基础上提出了基于树串映射和统计生成的基于实例的机器翻译方法[4]。

2.3.1 树串映射

如图 2 所示，树串映射（Tree-String Correspondence，简称 TSC）是一个三元组，包括：

- 源语言分析树；
- 目标语言串；
- 源语言树的叶结点和目标语言串之间的对应关系。

源语言分析树可以使用源语言分析器来获得，树上的每一个结点都用词和属性来标记。源语言和目标语言的对应关系可以通过双语词对齐来获得。树串映射既被用来表示经过预处理的静态翻译实例，又在翻译过程中被用来表示与输入句子相匹配的动态实例片断。

2.3.2 树串映射-树匹配

在翻译时，首先要找到与输入句子的分析树相匹配的 TSC。为了在多个可能的匹配中作出选择，我们使用了如下准则：

- 结点数较多的 TSC 优先；
- 语义相似度大的优先。

由于一个输入句子的翻译往往要通过组合多个实例的不同片段来完成，因此与输入句子相匹配的往往也不是单一的 TSC，而是多个 TSC 组成的 TSC 森林（TSC forest）。我们使用贪心算法来进行最佳 TSC 森林的搜索。

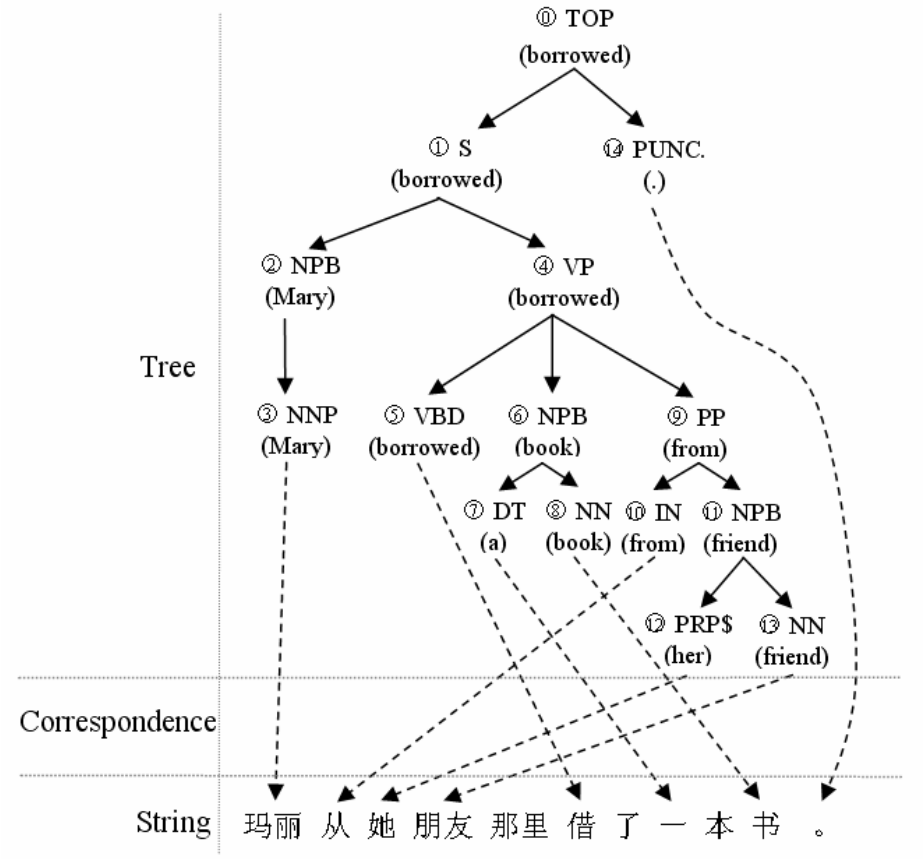

图 2　树串映射实例

2.3.3　统计生成

为了生成最终的译文，我们使用了统计生成模型来自底向上地组合 TSC 森林中的目标语言串。在统计生成模型中，使用了如下特征：

- TSC 和输入树之间的匹配程度；
- 源语言词翻译为目标语言词的翻译概率；
- 基于目标语语言模型的概率。

同时，为了生成更好的英文，我们利用未对齐词和同源 TSC（源语言树相同而目标语言串不同的 TSC）对翻译候选进行了扩展。

2.3.4　性能

用相同的训练集和测试集进行开放测试表明，基于树串映射和统计生成的基于实例的机器翻译系统的翻译质量好于基于词的统计机器翻译系统 ISI ReWrite decoder[15] 和基于短语的统计机器翻译系统 Pharaoh[16]。

2.4　统计机器翻译

在统计机器翻译模型中，词对齐模型是不可或缺的基础部分之一。同时，词对齐技术还还可应用于基于实例的机器翻译[4]、基于规则的机器翻译、计算机辅助翻译[17]、跨语言信息检索等各类系统中。因此，在过去两三年里，我们着重开展了词对齐技术的研究，其中多数研究是围绕着经典的 IBM 模型[18]展开的。统计机器翻译的其它方面及完整的统计机器翻译系统等研究目前正在进行中。

2.4.1　使用机器学习方法改进词对齐

我们首先感兴趣的问题之一是：在基本对齐模型和双语语料都固定不变的情况下，是否

能借助机器学习的方法来进一步提高对齐结果呢？为此，我们分别尝试了 bagging[19]、boosting[20]、semi-supervised boosting[21] 等方法。本质上，bagging、boosting 等方法都属于 ensemble 方法。Ensemble 方法通过组合多个分类器的决策来进行分类，从而得到比单个分类器更好的结果。

Bagging 方法随机地可重复地采样原始训练集来构造多个不同的训练集，从而训练得到不同的对齐模型。执行对齐过程时，首先可以根据这些对齐模型得到多个候选对齐结果，然后再通过加权的或不加权的多数表决方式来决定最终的对齐结果。

使用 Boosting 方法时，我们通过改变训练数据的权值来重新采样训练数据。其训练过程是一个多次迭代的过程。在每一轮迭代中，首先使用当前的加权训练集来训练对齐模型、然后使用得到的模型进行对齐，再根据对齐结果来计算错误率，最后根据错误率来为下一轮迭代重新计算权值。为了将 boosting 方法应用到无指导的词对齐中，需要解决参考集和错误率问题。对于参考集问题，我们用双向词对齐结果的交集和过滤过的并集来得到准确率较高的伪参考集。对于错误率问题，我们人工标注了少量训练语料作为开发集来计算错误率。

人工对齐大量双语语料是一件费时费力的工作，但加工少量语料还是相对容易的。使用少量人工对齐语料和大量未对齐语料，我们就可以使用 Semi-supervised boosting 的方法来提高词对齐了。在 boosting 过程中，我们将有指导训练得到的对齐模型和无指导训练得到的对齐模型进行线性插值得到一个统一的模型。

以上方法相对文献[22]中的 refine 方法分别取得了不同程度的提高，最好的 Semi-supervised boosting 方法相对于 refine 方法错误率降低了约 30%。

2.4.2 双语语料稀缺的语言对的词对齐

现在主流的词对齐技术通常需要较大的双语语料来训练对齐模型。而对于很多语言对，却很难获得足够多的双语语料。事实上，很多语言都有较多和英语之间的双语语料（例如汉英、日英），但与其它语言之间的双语语料却少得多（例如汉日）。为此，我们专门研究了如何使用有较多双语语料的两个语言对来帮助双语资源较少的语言对的词对齐[23]。以汉日词对齐为例，我们把英语作为枢轴语言，首先使用相对较大的汉英、日英语料来分别训练汉英、日英的词对齐模型，然后再根据汉英、日英的词对齐模型推导出汉日词对齐模型。这样推导出的汉日词对齐模型甚至可以达到比直接用小规模汉日语料训练出的模型还好的性能。当然，我们还可以将推导出的模型与直接训练的模型相结合，从而得到更好的性能。

2.4.3 特定领域的词对齐

即使对于有较多双语语料的语言对，如果把应用限定在特定领域里，语料仍然可能很少。为此，我们用较大的通用双语语料和很小的领域双语语料来分别训练模型，然后再将两个模型相结合得到更好的对齐结果。我们分别在结果上和模型上作了结合，对齐结果都得到了显著提高[24, 25]。

2.5 其它相关研究

机器翻译研究不是孤立的，它往往会涉及到自然语言处理的很多方面。为此我们也在中文分词、中文分析、中文语言模型、词义消歧、跨语言信息检索等方面开展了相关研究[26, 27, 28]。

2.6 多种方法的融合

2.6.1 系统级集成

对于不同机器翻译方法的结合，最直接也是最简单的方法就是制定一定的打分策略，然后直接在多个翻译结果中选择一个得分最高的作为最终的翻译结果。我们也正在尝试对基于

规则的系统和基于实例的系统的翻译结果进行选择，从而得到更好的总体性能。

2.6.2 模块级结合

与系统级集成相比，模块级结合要灵活多样得多。例如，在基于实例的系统中，我们就综合使用了多种技术：基于规则的分析器被用来进行源语言的分析、统计词对齐模块被用来建立源语言和目标语言间的对应关系、统计语言模型被用来对目标语建模、基于规则的后处理模块被用来做最终的生成处理。再如，我们用基于规则的机器翻译系统来帮助统计词对齐，使对齐结果得到了显著提高[29]。

3 讨论

基于规则的、基于实例的、以及统计的机器翻译方法是目前公认的对机器翻译方法的分类方式。通常认为，统计方法和实例方法都是数据驱动的方法。统计机器翻译方法在训练阶段使用双语语料获得统计模型，翻译时使用事先训练好的统计模型。基于实例的机器翻译方法本质上是一种基于实例推理的方法，它在翻译阶段直接使用翻译实例。与数据驱动的方法相对，传统的基于规则的机器翻译方法中，通常用人工撰写的规则来描述语言规律。

那么，三种机器翻译方法之间到底是什么关系呢？本质上，这三种方法并不是互斥的，而是着眼点不同，从不同角度强调问题的不同方面。统计是从数学角度强调统计建模能力，实例是从机器学习的角度强调对翻译实例的抽象程度，而规则是从语言现象入手来描述语言的成分构成规律。严谨的统计模型、大单元的翻译实例、多层次的成分构成完全是可以融合共生的。

4 结束语

统计机器翻译方法正在受到学术界更多的关注，基于规则的和基于实例的机器翻译方法也仍然有其重要价值。本文全面介绍了东芝对规则、实例、统计等各种机器翻译方法的研究，以及在多种方法相融合方面的探索。今后，我们将继续研究更有效的多种方法相融合的机器翻译策略。

参 考 文 献

1 Jiang Zhu, Haifeng Wang. 2005. The Effect of Adding Rules into the Rule-based MT System. In Proceedings of MT SUMMIT X, pages 298-304. September, 2005. Phuket Island, Thailand.

2 Andy Way, Nano Gough. 2005. Comparing Example-Based and Statistical Machine Translation. Natural Language Engineering, 11(3): 295-309.

3 Declan Groves, Andy Way. 2005. Hybrid Example-Based SMT: the Best of Both Worlds?. In Proceedings of ACL 2005 Workshop on Building and Using Parallel Texts: Data-Driven Machine Translation and Beyond, pages 183-190. June, 2005. Ann Arbor, MI.

4 Zhanyi Liu, Haifeng Wang, Hua Wu. 2006. Example-Based Machine Translation Based on Tree-String Correspondence and Statistical Generation. Machine Translation: Special Issue on Example-Based Machine Translation (in press).

5 Dekai Wu. 1997. Stochastic Inversion Transduction Grammars and Bilingual Parsing of Parallel Corpora. Computational Linguistics, 23(3):377-403.

6 Hiyan Alshawi, Srinivas Bangalore, and Shona Douglas. 2000. Learning Dependency Translation Models as Collections of Finite-state Head Transducers. Computational Linguistics, 26(1):45-60.

7 Dan Melamed. 2004. Statistical Machine Translation by Parsing. In Proceedings of ACL 2004, pages

653-660. July, 2004. Barcelona, Spain.

8 David Chiang. 2005. A hierarchical Phrase-based Model for Statistical Machine Translation. In Proceedings of ACL 2005, pages 263-270. June, 2005. Ann Arbor, MI.

9 Declan Groves, Andy Way. 2006. Hybrid Data-Driven Models of Machine Translation. Machine Translation: Special Issue on Example-Based Machine Translation (in press).

10 Shin-ya Amano, Hideki Hirakawa, Hirosysu Nogami, and Akira Kumano. 1989. The Toshiba Machine Translation system. Future Computing System, 2(3):227-246.

11 Arjen Poutsma. 2000. Data-oriented Translation. In Proceedings of COLING 2000, pages 635-641. Aug. 2000. Saarbrucken, Germany.

12 Andy Way. 2003. Machine Translation Using LFG-DOP. In Bod R, Scha R, Sima'anK(eds) Data-oriented parsing, pages 359-384. CSLI Publications, Stanford, CA.

13 Eiji Aramaki, Sadao Kurohashi. 2004. Example-Based Machine Translation Using Structural Translation Examples. In International Workshop on Spoken Language Translation, pages 91-94. September 2004. Kyoto, Japan.

14 Stuart M. Shieber. 1994. Restricting the Weak Generative Capacity of Synchronous Tree Adjoining Grammar. Computational Intelligence 10(4): 371-385.

15 Ulrich Germann. (2003) Greedy Decoding for Statistical Machine Translation in almost Linear Time. In Proceedings of HLT-NAACL 2003, pages 72-79. May, 2003. Edmonton, Alta, Canada.

16 Philipp Koehn. 2004. Pharaoh: A Beam Search Decoder for Phrase-Based Statistical Machine Translation Models. In Proceedings of Sixth AMTA, pages 115-124. September 2004. Washington DC.

17 Hua Wu, Haifeng Wang, Zhanyi Liu, Kai Tang. 2005. Improving Translation Memory with Word Alignment Information. In Proceedings of MT SUMMIT X, pages 313-320. September, 2005. Phuket Island, Thailand.

18 Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra and Robert L. Mercer 1993. The Mathematics of Statistical Machine Translation: Parameter Estimation. Computational Linguistics, 19(2):263-311.

19 Hua Wu, Haifeng Wang. 2005. Improving Statistical Word Alignment with Ensemble Methods. In Proceedings of IJCNLP-05, pages 462-473. October 2005. Jeju Island, Republic of Korea.

20 Hua Wu, Haifeng Wang. 2005. Boosting Statistical Word Alignment. In Proceedings of MT SUMMIT X, pages 364-371. September, 2005. Phuket Island, Thailand.

21 Hua Wu, Haifeng Wang, Zhanyi Liu. 2006. Boosting Statistical Word Alignment Using Labeled and Unlabeled Data. In Proceedings of Coling/ACL2006 Main Conference Poster Sessions, pages 913-920. July, 2006. Sydney, Australia.

22 Franz Josef Och and Hermann Ney 2000. Improved Statistical Alignment Models. In Proceedings of the 38th AC L, pages 440-447. September 2000. Hong Kong.

23 Haifeng Wang, Hua Wu, Zhanyi Liu. 2006. Word Alignment for Languages with Scarce Resources Using Bilingual Corpora of Other Language Pairs. In Proceedings of Coling/ACL2006 Main Conference Poster Sessions, pages 874-881. July, 2006. Sydney, Australia.

24 Hua Wu, Haifeng Wang. 2004. Improving Domain-Specific Word Alignment for Computer Assisted Translation. In Companion Volume to the Proceedings of ACL-04, pages 97-100. July, 2004. Barcelona, Spain.

25 Hua Wu, Haifeng Wang, Zhanyi Liu. 2005. Alignment Model Adaptation for Domain-Specific Word Alignment. In Proceedings of ACL-05, pages 467-474. June, 2005. Ann Arbor, Michigan.

26 Jianfeng Li, Haifeng Wang, Dengjun Ren, Guohua Li. 2006. Discriminative Pruning of Language Models for Chinese Word Segmentation. In Proceedings of Coling/ACL2006, pages pages1001-1008. July, 2006.

Sydney, Australia.

27 Zhimao Lu, Haifeng Wang, Jianmin Yao, Ting Liu, Sheng Li. 2006. An Equivalent Pseudoword for Unsupervised Chinese Word Sense Disambiguation. In Proceedings of Coling/ACL2006, pages 457-464. July, 2006. Sydney, Australia.

28 Jiang Zhu, Haifeng Wang. 2006. The Effect of Translation Quality in MT-Based Cross-Language Information Retrieval. In Proceedings of Coling/ACL2006, pages 593-600. July, 2006. Sydney, Australia.

29 Hua Wu, Haifeng Wang. 2004. Improving Statistical Word Alignment with a Rule-Based Machine Translation System. In Proceedings of COLING-04, pages 29-35. August, 2004. Geneva, Switzerland.