机器翻译中汉语词节点的识别[1]

王厚峰
北京大学信息科学技术学院，计算语言学研究所，100871
wanghf@pku.edu.cn

摘 要：汉语词节点的识别包括纯切分歧义消解、命名实体识别和缩略语的辨识等，被认为是中文信息处理的困难问题。为了考察难点所在，我们测试了两个机器翻译系统识别词节点的性能。结果表明，命名实体和缩略语识别错误是词节点识别错误的主要来源。本文对此作了评估和讨论。

关键词：分词，命名实体识别，缩略语辨识

Evaluation on Chinese Tokenization in Machine Translation

Wang Houfeng
Institute of Computational Linguistics, Peking University, Beijing, 100871
wanghf@pku.edu.cn

Abstract: Chinese Tokenization consists of pure word segmentation disambiguation, named entities recognition and abbreviation identification etc. By evaluating tokenization performance of two MTs, we found that errors rate of named entities and abbreviation identification is much higher than that of pure Chinese word segmentation disambiguation. In this paper, some test data is given and relative issues are discussed.

Keyword: Chinese Word Segmentation, Named Entity Recognition, Abbreviation Identification.

1. 引言

词节点识别，即分词，算得上是中文信息处理的传统问题。它不仅吸引了国内许多研究人员的长期重视[孙茂松，邹家彦，2001]，而且，近年来，该问题也引起了国外一些研究人员的关注[Hockenmaier, et al ,1998 和 Teahan, et.al,2000]。

词节点识别被认为是中文信息处理的困难问题。已有很多文章对此进行评述。本文则希望通过对测试数据的分析，进一步对问题进行分类。我们选择的测试系统分别是华建和 Babelfish 的网上即时翻译（本文用 MT-A 和 MT-B 表示）。测试语料选自于 LDC 的 105 篇新闻稿，其中，新华社文稿 52 篇、美国之音文稿 26 篇，新加坡联合早报文稿 27 篇。

为了便于分类讨论，我们先将词节点识别中所涉及到的词作如下划分：

[1] 本文受国家自然科学基金资助，编号：60173005

（1） 已知词。分词词典中收录的词。

（2） 可构造性词。没被分词词典收录，但可以给出构词规则，如数量词。

（3） 部分可构造性未登录词。这类词有一些构成上的特征，但不能给出完整而明确的构词规则，如命名实体，简称缩略语等。它们数量大，变化快。

（4） 没有任何构成特征的未登录词，即，真正的新词。如“克隆”，“纳米”。

本文只考察第（1）（3）类情况。一方面，第（2）类已较好解决，第（3）类需要深层分析，过于困难；另一方面，第（1）（3）类代表了词节点识别中的绝大部分情况。

2. 切分歧义消解

本文所指的切分歧义，是指不考虑上节所列的第（2）-（4）类未登录词情况下可能存在的歧义（歧义程度与词典大小密切相关），包括交叉性歧义和包容性歧义。

切分歧义消解方法可分为基于词典的机械式匹配和基于语料库的动态计算方法。前者主要包括正、逆向的最大长度匹配，后者则包括互信息方法，n-gram，PPM+HMM 模型和基于变换的错误驱动等方法，相关技术有很多文章报道。但无论是哪种方法，都无法完全消解两类歧义中的任何一类。先看两个例子：

例 2.1 (1) 以总成绩３５５・３５分居领先地位

MT-A: **live apart** the very latest with total 355·35 achievement

MT-B: by total result 355.35 **lives apart** the leading status

(2) 丁豪成为中国孤儿院长大的第一个残疾人大学生

MT-A: Hao Ding become **president** of Chinese orphan **heavy** first disabled person university student

MT-B: Ding boldly became Chinese orphan Chair big first deformed National People's Congress student

如果只考虑切分歧义，上例中的 (1) 有包容性歧义，存在“分居”与“分 / 居”两种分词。上例 (2) 有多处歧义，如“院长大”和“残疾人大学生”。其中“院长大”存在交叉歧义，可以切分成“院长 / 大”和“院 / 长大”。

纯粹切分歧义错误的数据如下表：

Table 2.1: Number of word segmentation errors

	新华社文稿	美国之音文稿	新加坡联合早报
MT-A	5	2	3
MT-B	6	8	6

显然，纯切分歧义的错误并不十分严重。但一个分词错误是否是纯粹的切分歧义错误，其判断可能存在主观因素，因为我们并不完全清楚哪些词已经在词典中。如：

例 2.2 泰国财长塔林：Thai **wealth long** Tallin

“财长”被切分成了“财 / 长”，甚至“长”也选择了错误的发音。但“财长”可能属于可构造性未登录词（“财政部长”的缩写）。因此，该类错误未计入上面的表中。

3. 命名实体的识别

部分可构造性的词包括命名实体和简称缩略语。这类词在新闻语料中大量出现，成为词节点识别中的突出问题。命名实体包括人名，地名，机构名，商标名和货币名等，本文主要考察人名、地名和机构名。

在人名、地名和机构名三类命名实体中，人名的指示信息最为丰富，可以直接在前后面出现称呼、称谓、头衔和人所特有的属性与动作等。先看三个错误的例子：

例 3.1 (1) 张旭明表示

MT-A: **Zhang Xu** will express **tomorrow**

MT-B: **Zhang Xu** clearly indicates

(2) 600 多万张伟哥处方

MT-A: more than 6 million **Zhang Wei's brother's** prescription

MT-B: more than 6000000 **Zhang Wei Ge** prescription

(3) 加拿大总理科雷迪安 (新加坡联合早报)

MT-A: Canadian Premier **Dean in Kley**

MT-B: the Canadian **total science subjects 雷迪 where**

如果只考虑名字前后紧相邻的直接指示信息，则在测试的总共 461 个（包括重复出现）人名中，有 297 个人名带有直接指示信息，占 297/461=64.43%，再看人名识别错误的情况：

Table 3.1: 全部人名的错误情况

Xinhua +VOA+ZBN	全部人名数目	错误识别数目	错误率
MT-A	461	192	41.6%
MT-B	461	311	67.5%

与上一节纯粹切分歧义的错误相比，不仅其数目大，而且其错误率也高。即使有直接指示信息，两个系统错误率仍远超出了 (1 – 64.43 %) = 35.57% ，因此，如何利用指示信息，是一个需要进一步研究的问题，例 3.1 (1) 中的“表示”和(3) 中的“总理”都是非常明确的指示信息，但是，有干扰，主要是长度干扰和用字干扰。另一方面，如何有效地利用“姓”的特征，也值得深入研究，如果处理不当，有可能会将其功能扩大化。例 3.1 (2)就将表示量词的“张”当成了“姓”，从而导致错误。

地名和机构名也是两类重要的命名实体。它们的指示信息相对较弱．主要出现在命名实体后面，如北京市的“市”，联想集团的“集团”等。

下面是与地名和组织机构名相关的例子：

例 3.2

(1) 福州**鼓山镇**的**福兴**投资区、**晋江 安海**的**桥头**工业区均成为全国乡镇企业示范小区

参考译文: The **Fuxing** Investment Zone in **Gushan Town** of **Fuzhou** and the **Qiaotou** Industrial Area in **Anhai** of **Jinjiang** have both become the country's model communities for township enterprises.

MT-A: Mountain, drum of Foochow, good fortune of town revitalize person who make the

investment, Jinjiang either end of a bridge industrial area to set up sea become the national demonstration district of township enterprise .

MT-B: The Fuzhou drum Shan Zhen Thualuu investment area, the Jinjiang An Hai bridge head industrial district all becomes the national rural enterprise demonstration plot.

例 3.3 (1) 连云港**如意**集团

MT-A: The group **as one wishe**s of Lianyun Harbour

MT-B: Lienyungang pleasant group

(2) 由美国**永道**会计师事务所具体承办

MT-A: **Said** the accounting firm to undertake concretly **forever** by U.S.A.

MT-B: by USA **forever will say** accountant the office specifically will undertake

由于很多地名(如国家，首都等)可能被收录进词典，无法界定哪些地名“未知”，因此，很难给出反映真正未知地名识别的错误率。但从我们的感觉来看，没有收进词典的，错误率就很高。例 3.2 中如果不考虑“福州”，则只有“晋江”识别正确（其实，我们也无法判断“晋江”是否被收录到词典中）。

但对于机构名，词典收录则要少得多。因为机构名并不像地名那样相对稳定，时常有旧机构消失和新机构产生，有的机构还会根据需要改名称。例 3.3 是机构名识别错误的两个例子；下表给出了机构名识别错误的统计数据：

表 3.2 机构名的测试结果

	总数	带指示信息	错误数
MT-A	61	49	34
MT-B	61	49	42

机构名的指示信息主要有“公司”，“部”，“委”等。从出错的情况来看，两个系统并未有效运用这些信息。

从语义上讲，机构名介于人名和地名之间（具有人的功能，通常也有地理上的意义）；但从名称本身的构成上讲，机构名与地名有更多的相似性，如，引导信息都在后面（北京**市**，联想**集团**），长度上也都比较自由（如喜马拉雅山，中国电气进出口联营公司）；但机构名称也有特别之处：大量的机构名可以切分成多个普通词，如“如意/集团”，这就存在两个问题：其一是，如何判定其是机构名，还是普通词序列，如，“北京/大学” 是机构名，而“香港/公司”则不是；其二，当机构名可以切分成多个词时，如“国家/科技/奖励/工作/办公室”，如何确定其边界词？

对于没有形式上的指示信息、也没有收录到词典中的未知地名和机构名的辨识，通常要作较为深层的分析，特别是要作语义分析，这里不详细讨论。但另一方面，形式标记可以进一步挖掘，如表示平行关系的标点符号逗号和顿号，特殊区域使用的标记，如“旗”常作为内蒙古的地名后缀。对于已经有指示信息的情况，特别是机构名识别，也存在较大的改进余地。例如，“香港公司”不能作为机构名，但“香港大学”则可以，这就需要利用不同指示信息的不同构词特点。同样，在右边界明确的情况下，其左边界用词也有特点：它们常常会是地名，或者某些特殊的词，如“国家，国际，世界”等。

4. 简称缩略语识别

另一类部分可构造性词是简称缩略语。之所以归入部分可构造类，主要因为大部分缩略语中的字顺次来自于全称。简称缩略语既包括简称，又包括缩略语。

简称通常针对命名实体，如“北大”常作为“北京大学”的简称；而缩略语则常指命名实体之外的词串的压缩。如“低保”作为“最低生活保障”的缩略语。本文对两种情况不作区别。在汉语上，还因为特殊的历史、文化等因素，一些简称与全称在用字上完全没有关联，如“沪宁高速公路”中的“沪”表示“上海”，“宁”表示“南京”。这类简称属于别名式简称，不属于部分可构造类。因此，不是本文讨论的范畴。

缩略语有很多研究，但在中文信息处理中，如何自动识别缩略语，还没有看到报道。

简称缩略语在被广泛接受后，会变成常规词。如“中国”是“中华人民共和国”的简称，但已经成为常规词。此外，简称也是相对的，如“中”又是“中国”的简称。

同全称相比，汉语中的简称缩略语有多种表现形式：

(1) 如果原有全称是一个词，则其中的首字或尾字可以作为简称，如，中国一中，澳门一澳，香港一港；但首字的使用更为常见；

(2) 如果原有全称由多个词构成，则，可以依次取几个词的首字(类似于 Acronym)，如“北京 / 大学”简称为“北大”；也可以直接取第一个词，如“清华 / 大学”简称为“清华”；有时也直接取全称的首尾字，如“财政部长”与“财长”；当最后一个字具有语素作用时，常常会在简称缩略语中出现，如“中国科学院”与“中科院”，“对外经济贸易部”与“外贸部”。

如下是从测试中取出的例子，这些例子在两个机器翻译中均有错误：

例 4.1

(1) **日相希望宫泽**留任(日本首相希望宫泽喜一留任，例子源于《联合早报》)

MT-A：Japan hopes that the **pool** retains the office in the **palace**

MT-B：**date** hopes **palace Ze** remains in office

(2) 为**中吉两国**进一步合作...，

MT-A：For **in two lucky country** cooperate in further

MT-B：further cooperated for **center lucky** both countries

下表给出了简称缩略语识别的错误统计情况，我们认为可能已经出现在辞典上，而且很难引起歧义性的一部分没有计入表中（如英-英国，柬-柬埔寨等）。

表 4.1 简称缩略语的统计结果

	总数	错误数
MT-A	66	32
MT-B	66	48

简称缩略语的识别有如下两个特殊困难：

其一，如何判断某个字序列为简称缩略语。这又可以分为两个问题：首先，要与真正的新词区分；其次，一个简称缩略语可以对应多个全称，甚至可切分成普通词序列，如“中

巴”可能代表“中国-巴基斯坦”、“中国-巴西”或“中国-巴勒斯坦”等，还可以表示一种巴士。如何区分是一个问题。

其二，简称缩略语的识别受到相应的全称识别的影响。测试发现，简称缩略语主要来源于命名实体，而命名实体的正确率本身也是一个问题。

如果过于强调对简称缩略语的处理，也会导致错误。如下面的例子：

例 4.2

(1) 新项目中外商投资比例越来越高：new project **Chinese and foreign traders** are higher and higher in ratio between investments

(2) 姓闽的年轻人

MT-A：young man of **Fujian**

MT-B：surnames Fujian Province's young people

要解决上述问题之一，必须依赖上下文，寻找对应的全称。这一方面需要较好的匹配模式和快速匹配算法，同时，还需要优先保证全称的正常识别，特别是命名实体全称识别正确；另一方面，由于这种解析过程本质上是“共指消解”(Coreference Resolution)过程，还应该引入共指消解的技术，本文对此不作详细讨论。

5. 总结

关于词节点识别的研究，存在很多争议。比如，自然语言处理是否需要建立在分词基础上以及彻底分词的时机是什么。在中文信息处理的一些应用上，人们也的确在尝试着没有分词的处理方法，但在很多应用方面，特别是机器翻译，准确的词识别仍然是必须的。这就需要探讨分词究竟存在哪些问题和问题的根源何在。从我们测试的结果来看，命名实体和简称缩略语识别的准确度低是问题的关键，如果进一步排除可能收录进词典的命名实体（如江泽民，克林顿，诺罗敦.西哈努克），准确度将更低。我们也发现，命名实体和简称缩略语识别存在可提升的空间。这一方面可以通过充分挖掘和有效利用命名实体的指示信息改善命名实体识别的性能，同时，其性能改善也有助于提高命名实体简称的识别能力。此外，在简称缩略语的识别中，还可以通过探索和使用共指消解技术提高其准确性。

参 考 文 献

[1] 孙茂松. & 邹家彦. (2001) “中文自动分词研究评述”, 当代语言学，Vol.3, No.1, pp 22-32

[2] Hockenmaier, J. and Brew, C. (1998) “Error-driven Learning of Chinese word segmentation” In J. Guo, K.T. Lua and J. Xu editors, *12th Pacific Conference on Language and Information*, pp.218-229, Singapore.

[3] Teahan, W. J., Wen, Y., McNab, R., Witten, I. H.(2000) “A compression- based algorithm for Chinese word segmentation”. *Computational Linguistics*, Vol.26, No.3, pp. 375-393.