汉语块分析评测任务设计*

周强　李玉梅

清华大学信息技术研究院语音和语言技术中心

清华信息科学与技术国家实验室，北京 100084

E-mail: zq-lxd@mail.tsinghua.edu.cn, limy@cslt.riit.tsinghua.edu.cn

摘要：本文介绍了目前正在筹备的中文信息学会句法分析评测 CIPS-ParsEval-2009 中的三项块分析评测任务：基本块分析、功能块分析和事件描述小句识别的设计理念、判定标准和相关资源构建方法，并通过相关统计数据分析和国内外相关研究评述，总结了这三项评测任务的主要特色。

关键词：块分析评测，基本块，功能块，事件描述小句，块标注库

Chinese Chunk Parsing Evaluation Tasks

Qiang Zhou, Yumei Li

Center for Speech and Language Technology, Research Institute of Information Technology

Tsinghua National Laboratory of Information Science and Technology

Tsinghua University, Beijing 100084

E-mail: zq-lxd@mail.tsinghua.edu.cn, limy@cslt.riit.tsinghua.edu.cn

Abstract: The paper introduces three chunk parsing tasks proposed in current CIPS parsing evaluation workshop (CIPS-ParsEval-2009) that is organized by Tsinghua University and North-east University. They are base chunk parsing, functional chunk parsing and event description clause recognition tasks. The designing ideas and classification standards of these three chunks are discussed in the paper. Based on the detailed syntactic annotations in Tsinghua Chinese Treebank (TCT), three benchmark chunk banks automatically extracted from TCT are built. The data analysis from their statistics and the comparison with current different chunk schemes show some characteristics of these three chunk parsing tasks.

1. 引言

有效的真实文本评测任务设计是提升自然语言处理技术的一个重要途径。英语方面的一个典型例子 CoNLL 设计的一系列共享分析任务，包括基本名词短语识别[1]、文本块分析[2]、子句识别[3]、命名实体识别[4, 5]、语义角色标注[6, 7]、依存分析、句法依存和语义角色一体化处理等，从简单到复杂，通过设计合适的分析任务，构建共享评测数据(Benchmark)，吸引了国内外大量感兴趣的研究人员探索了各种机器学习模型在不同的分析任务中的应用方法，开发出一组可共享的英语文本句法语义分析工具。

在汉语方面，从 2003 年起，SigHan 分别组织了 3 届汉语词语切分评测 Bake-off，大大推动了相关研究技术的发展。2007-2008 年，又与中文信息学会联合举办了第 4 届 Bake-off 评测[8]，进一步增加了汉语词性标注和命名实体识别评测任务。但与英文相比，在句法语义分析层面上的

*本文工作得到了国家自然科学基金资助项目（编号：60573185, 60873173）和国家 863 计划资助课题（编号：2007AA01Z173）支持。

评测任务则比较少。

受中文信息学会委托，从 2008 年 10 月起，清华大学和东北大学开始筹办中文信息学会句法评测 CIPS-ParsEval-2009[9]。其主要目标是针对汉语描述特点，设计合适的评测任务，开发有效的评测数据集合。并以此为契机，推动国内汉语文本句法分析的研究水平。通过深入研究，我们提出了 5 项评测任务，其中 3 项涉及汉语文本的块分析问题。本文将对有关内容进行具体介绍和说明。

2. 块分析任务设计

我们从事件内容分析应用角度，把汉语句子的完整句法分析任务设计为事件描述小句识别、事件描述结构分析和事件逻辑关系分析三个阶段。考虑到这个分析任务的复杂性，在本次评测中，我们把研究重点集中在事件内容的句法结构分析方面，从中提炼形成了以下 3 项块分析子任务：1）事件描述小句识别；2）功能块分析；3）基本块分析。它们形成既相互独立、又互有联系的分层次的块分析描述体系。

我们的块分析体系设计的基本理念是：块是句法语义信息的结合体，块内部的词语关联性是句法语义联系的桥梁。一个理想的块设计应该既能找到明确的句法判据，又可以形成合理的语义解释，达到形式和意义的完美结合。目前，基本块主要采用了内聚性判据，通过分析其内部词语组成的不同拓扑结构特点来判断是否成块；功能块和事件描述小句主要采用了外延性判据，通过分析它们在更大的事件句式和复杂句子中所处的功能位置及其与其他相邻成分的句法语义关系来判断是否成块。下面几节将对有关内容进行简要说明。

1）基本块（Base Chunk, BC）

我们把基本块定义为单个或多个实词按照一定的关联关系组合形成的基本信息单元[11]。通过对基本块内部各种词汇关联关系的深入分析，我们提炼出了三种典型拓扑结构：左角中心结构(LCC)、右角中心结构(RCC)和链式关联结构(CHC)，它们覆盖了基本块内部的以下句法关联关系：1）修饰关系：覆盖体词块和谓词块 RCC 和 CHC；2）并列关系：覆盖体词块和谓词块 CHC；3）述宾、述补和附加关系：覆盖谓词块 LCC。

这样，就形成了以下基本块内聚性判据：1）句子中的实词组合符合上面的一种拓扑结构，则形成一个多词语基本块；2）句子中的其他独立出现的实词直接形成一个单词语基本块。对分析出的每个基本块，将给出“成分标记+关系标记”的双标记描述[11]。

2）功能块（Functional Chunk, FC）

汉语功能块主要描述句子中反映不同事件内容的基本单元。他们一般占据了句子中的主语、谓语、宾语、状语、定语、中心语等功能位置，通过组合形成不同的事件句式完成对真实世界的不同事件内容的再现描述。功能块的确定主要依据它们在事件描述小句的不同层次事件句式中所处的功能位置。目前主要考虑了以下两类事件句式：1）小句层面上的基本句式结构。据此，可以确定主、谓、状、宾、补等功能块。2）复杂名词短语层面上的句式结构变体。据此，确定定语块、中心块等功能块。

但是，真实文本的事件描述小句中的事件句式的层次关系是很复杂的。如：某个句式变体可以直接充当基本句式中的主、宾等成分；某个基本事件句式也可以直接作为一个整体充当某个小句事件句式中的主、宾、状等功能成分，形成主语、宾语或状语从句。这些事件句式可以组合

形成一个分层次的事件骨架树结构。为了简化起见，在本次评测中，我们只考虑各个事件描述小句的事件骨架树中最低层次（即叶子节点）的功能块，将它们按照从左到右的顺序排列形成整个事件描述小句的功能块标注序列。

这样，就形成了以下功能块外延性判据：选择事件描述小句的事件骨架树中最低层次（即叶子节点）的词语组合形成各个功能块。对分析出的每个功能块，将分别使用以下 10 个功能标记来标注：主语块（S）、状语块（D）、述语块（P）、宾语块（O）、补语块（C）、兼语块（J）、定语块（A）、中心块（H）、独立块（T）和其他特殊块（X）。

3）事件描述小句（Event Descriptive Clause, EDC）

我们以句号、问号和叹号等作为完整汉语句子的分隔符。在此基础上的事件描述小句确定主要依据了以下判定条件：1）以逗号、分号、句号、问号等点号分隔而形成的词语序列；2）内部包含完整的主、状、谓、宾等事件句式，考虑到各种省略情况，其中至少应包含一个谓语块；3）复句层面的状语和独立语成分可以作为一个特殊的 EDC。它们共同形成 EDC 的外延性判据。

我们使用以下 4 个标记来标注不同的 EDC：1) E1--包含主题信息的 EDC；2）E2--主题信息省略的 EDC；3) D1--复句层面的状语块；4) T--复句层面的独立语块。其中 E1 和 E2 组成了典型的事件描述小句。

3. 评测数据库开发

以 TCT 作为统一的数据源，充分利用其中提供的丰富句法成分和关系标记信息，将上面设计的三种块的句法判据进行具体化和实例化，我们可以自动提取形成不同的块标注语料库，从而可以对这三个不同层次的块分析任务的处理难度进行初步估计。在下面实验中，主要选择了 TCT 中所有的新闻类文本。其基本统计数据是：文件数 185，汉字总数 325806，词语项总数 207372，句子总数 8137，平均长度为 25.49 词/句。

1） 基本块数据分析

从 6 个主要基本块的长度分布数据可以看出[11]，真实文本句子中描述实体内容的名词基本块和描述动作状态的动词基本块占了大多数，达到单词语块总数的 91%和多词语块总数 77%，是我们研究的重点。相对而言，动词块的平均长度较短。在多词语块中，只包含 2 个词语的块占了 93%以上；而在 np 多词语块中，包含 2 个词语的块只占了 71%左右，约 30%的名词块长度超过了 3 个词语。因此，基本名词块的内部描述复杂度更高，进行自动准确分析的难度也更大。

表 1 功能块长度分布

功能块	块总数	词总数	平均长度
P-谓语	31213	45558	1.4596
D-状语	18780	37889	2.0175
S-主语	13531	31441	2.3236
O-宾语	11550	42258	3.6587
H-中心	4804	9784	2.0366
A-定语	1125	2615	2.3244
J-兼语	1042	2355	2.2601
T-独立	372	790	2.1237
C-补语	282	535	1.8972
X-其他	237	1026	4.3291

2） 功能块数据分析

表 1 列出了功能块长度分布数据。从中我们可以发现：

- 真实文本句子中 P、D、S、O 块占了绝大多数，它们是形成事件句式的基本单元。其中的主要识别难点是复杂的宾语、状语和主语块。而 P 块一般都由基本 vp 和 ap

块充当，自动识别难度相对较小。

- H 和 A 块主要出现在定语从句中，其平均长度和分布特点基本与 S 块相当，但由于出现数量较少，再加上汉语典型歧义结构“V N 的 N”的影响，会导致统计学习模型训练不充分，从而增大识别难度。而 H 块由于前面一般有助词‘的’，会更容易识别。
- 在剩余的 4 个非典型功能块中，J 和 C 尽管出现频度较少，但由于语境特征明显，其识别难度应该与 H 块相当。而 T 和 X 则由于组合情况复杂和语境分布特征不明显，自动识别难度会很大，但由于其绝对数量很少，对整体性能的影响可以忽略。

综上所述，在我们关注的 8 个功能块（PDSOHAJC）中，预期的识别难度排列会是：P, 简单 D,S,O < H, J, C < 复杂 D, S, O < A。另外，由于 92%以上的基本块和功能块可以方便地建立直接联系（1:1 + N:1(S)），而且的它们的标注信息又有很强的信息互补性，通过适当的融合处理，我们可以方便地得到信息更完整的功能块（功能标记+成分标记+中心词位置），从而为下一步的事件句式识别和事件骨架树分析打下很好的基础。

3） 事件描述小句数据分析

表 2 列出了不同类型的事件描述小句的长度分布数据。图 1 显示了其中不同长度 EDC 的分布比例。从这些数据可以看出：

表 2 事件描述小句长度分布

EDC	块总数	词总数	平均长度
E1	11282	119104	10.557
E2	7870	62057	7.885
D1	1167	5006	4.290
T	110	263	2.391
合计	20429	186430	9.126

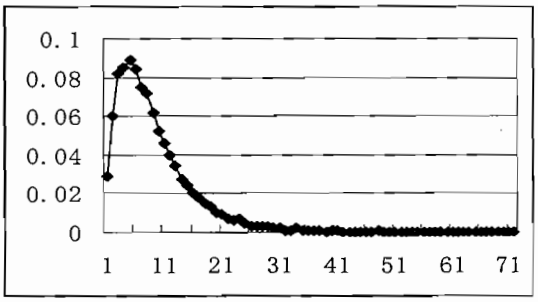

图 1 不同长度 EDC 所占比例分布

- 汉语真实文本中包含完整事件内容的典型 EDC 块(E1+E2 类)占了 95%以上，是自动识别研究的主体。
- 典型 EDC 块的平均长度达到 9 个词以上，远高于功能块和基本块，并且长度大于 10 个词的 EDC 块比例超过了 30%，长度大于 20 个词的 EDC 块比例也达到了 6%，这就进一步加大了相关 EDC 块的识别难度。
- 点号作为事件描述小句的天然分隔符，应该可以在 EDC 识别中发挥重要作用。但汉语点号使用非常灵活，可用来分隔主、状、宾等功能块，可用来分隔各个功能块内部的并列成分，也可用来分隔复杂从句内部的各个小句，以上这些情况在我们目前的 EDC 划分原则下都应包含在某个 EDC 内部。对目前的 2 万多个 EDC 进行内部信息分析，发现包含点号的 EDC 占块总数的 16%，占覆盖词语总数的 32%。这表明仅仅依靠点号信息来切分 EDC 会带来很大的副作用，需要引入更多有效的判别特征。
- 汉语事件描述小句内部的功能块组合非常复杂，包含多个谓语块的 EDC 比例达到了 37%以上，其中包括复杂从句和连谓、兼语、并列等复杂谓语结构，它们会形成复杂的事件句式和事件骨架树。这些情况与灵活的点号使用习惯混杂在一起，对准确识别表征完整事件描述内容的 EDC 任务，提出了很大的挑战。

4. 相关研究工作评述

在基本块层面，英语方面的工作主要基于 Abney(1991)提出的语块(chunk)概念[18]。它被定义为句子中属于同一个 S-投射的相邻词语所组成的词语串，从而建立了语块与管辖约束理论的 X-bar 系统的内在联系。CoNLL-2000 在华尔街日报语料库上进行的全面测试表明，在这个体系下建立的英语基本名词和动词块的识别性能达到 93%左右[2]。在汉语方面的类似工作有清华[13]和哈工大[14]的基本短语描述体系和微软的块描述体系[15]等。这些体系的共同点在于它们都是从句法层面上来定义和描述块信息，主要侧重块边界确定和句法成分标注问题，不太关心各个块的内部关系分析。另一类相关的研究则关注类似基本块的实词组合的整体语义表现和内部组合关系，典型的工作包括命名实体定义和识别[4,5]、多词表达的内部词汇语义组合性评估问题[12]等。

而我们提出的基本块描述体系则以语义中心驱动的典型拓扑结构分析为基本判据，将以上两部分的工作有机结合起来，达到了基本块形式和意义的初步融合。另外，还首次将紧密结合的述宾结构关系纳入基本块描述体系中，使之基本覆盖了汉语中所有实词之间的重要词汇关联关系，为在此层面上进行汉语词汇关系的自动获取研究打下了很好的基础。

在功能块层面，英语方面的研究主要集中在语义角色标注(SRL)方面，通过对句子进行浅层语义分析，确定各个目标动词控制的核心语义角色的准确边界，在语义层面上直接完成事件框架的分析识别。目前在英语 Propbank 测试库上的最好系统的 SRL 性能 F 值达到了 80%左右[7]，近几年也没有很大性能提升[19]。对实验结果的深入分析发现，其中的主要问题出在论元成分识别阶段：在 81%边界识别正确的论元成分中，95%以上都可以准确标注上合适的语义角色[7]。而且核心角色和外围角色的识别性能差异明显（80% VS 60%），显示出一定的统计偏置性。

而我们的研究则侧重从句法层面先识别出进行可以充当论元成分的功能块以及相应的事件句式，从而抓住了 SRL 的核心问题。这个研究从最初的单层次功能块[16]，到逐步细化的二层次功能块[17]，到目前的覆盖所有基本事件描述小句的功能块，再配合以事件骨架树的准确分析，可以实现语义层面的 SRL 在句法层面上的有效模拟。

在事件描述小句层面，国内外的相关研究不是很多。CoNLL-2001 曾提出一个英语子句识别任务[3]，其目标是自动识别英语句子中的所有嵌套子句。考虑到这个问题的复杂性，他们把它拆分成三项子任务：子句起点识别、终点识别和完整嵌套结构识别。其中最困难的第三项子任务基本上与我们定义的事件描述小句识别任务相当。只是我们只处理最上层的 EDC。当时最好系统的开放测试 F1 值为 78.63%[3]，后来，通过改进算法，将分析性能提高到了 80.44%[20]。

英语子句一般由先行词引导，具有比较明显的形式标记。这是设计嵌套子句识别任务的描述基础。而汉语各个从句之间一般没有特别的形式标记，因此我们选择以点号分隔的 EDC 作为识别重点，可能更适合汉语的描述特点。

5. 总结与展望

本文针对汉语的描述特点，提出了三项汉语块分析评测任务：基本块分析，功能块分析和事件描述小句识别。基于真实文本标注库的数据统计分析和国内外相关体系的对比分析研究显示，这套块分析评测任务设计具有以下特点：1）在基本块层面，以语义中心驱动的拓扑结构分析作为基本块的主要判据，并加入紧密结合的述宾关系描述，使之基本覆盖了汉语中所有实词之间的重要词汇关联关系；2）在功能块层面，选择不同层次事件句式中的各个最小描述单元作为

处理对象，最大限度地保留了句子中各个不同层面的事件描述信息，形成了进行事件骨架树分析的研究基础；3）在事件描述小句层面，以点号分隔的完整事件单元识别作为突破口，可以形成进行汉语“句→段”意合分析的中枢桥梁。

在此基础上，下一步的研究方向是：1）利用基本块和功能块的信息互补特点，通过适当的融合处理，获取信息更完整的功能块（功能标记+成分标记+中心词位置），以此作为事件骨架树分析的叶子节点；2）探索有效的事件骨架树分析方法，准确识别句子中由功能块组合形成的不同层次的事件句式，补充“功能块→事件描述小句”之间的事件信息描述空白。

参考文献

[1] Introduction to CoNLL-1999 Shared Task: NP braketing, http://www.cnts.ua.ac.be/conll99/

[2] Erik F. Tjong Kim Sang and Sabine Buchholz. (2000). “Introduction to CoNLL-2000 Shared Task: Chunking” [A]. *Proceedings of CoNLL-2000 and LLL-2000.* [C] Lisbon, Portugal. 127-132.

[3] Sang T K and Déjean H.(2001). Introduction to the CoNLL-2001 Shared Task: Clause Identification [A]. In *Proc. of CoNLL-2001* [C], Toulouse, France, p53-57.

[4] Erik F. Tjong Kim Sang (2002) Introduction to the CoNLL-2002 Shared Task: Language Independent Named Entity Recognition[A]. In *Proc. of CoNLL-2002* [C]

[5] Erik F. Tjong Kim Sang & Fien De Meulder (2003) Introduction to the CoNLL-2003 Shared Task: Language Independent Named Entity Recognition[A]. In *Proc. of CoNLL-2003* [C]

[6] Carreras, X. and M`arquez, L. (2004). Introduction to the conll-2004 shared tasks: Semantic role labeling [A]. In Proc. of CoNLL-2004[C].

[7] Carreras X. and M`arquez, L. (2005). Introduction to the conll-2005 shared tasks: Semantic role labeling [A]. In Proc. of CoNLL-2005[C]

[8] Guangjin Jin, Xiao Chen (2008) The Fourth International Chinese Language Processing Bakeoff: Chinese Word Segmentation, Named Entity Recognition and Chinese POS Tagging [A]. In Proc. of *Sixth SIGHAN Workshop on Chinese Language Processing*[C].

[9] 中文信息学会句法分析评测 CIPS-ParsEval-2009 介绍. http://www.ncmmsc.org/CIPS-ParsEval-2009/.

[10] 周强（2004） 汉语句法树库标注体系 [J]，《中文信息学报》，18(4), p1-8.

[11] 周强（2007）汉语基本块描述体系[J]. 《中文信息学报》，21(3)，p21-27.

[12] Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger (2002) “Multiword Expressions: A Pain in the Neck for NLP” [A]. In Proc. Third International Conference of Computational Linguistics and Intelligent Text Processing (CICLing 2002) [C], Mexico City, Mexico, February 17-23, 2002.

[13] 张昱琪，周强 (2002). 汉语基本短语的自动识别 [J]，《中文信息学报》，16(6)，1-8.

[14] Tiejun Zhao, Muyun Yang et al.(2000) “Statistics Based Hybrid Approach to Chinese Base Phrase Identification” [A]. Proc. of the Second Chinese Language Processing [C]. ACL 2000, Hong Kong.

[15] Li, H., C. N. Huang, J. Gao, and X. Fan (2004) “Chinese Chunking with Another Type of Spec” [A]. In *Proceedings of the 3rd ACL SIGHAN Workshop*[C], Barcelona, Spain, 2004, pp. 41-48.

[16] 周强，赵颖泽（2007）汉语功能块自动分析 [J]. 《中文信息学报》，21(5), p18-27

[17] 陈亿、周强、宇航（2008）分层次的汉语功能块描述库构建分析 [J]. 《中文信息学报》. 22(3), p24-31.

[18] Steven Abney(1991). “Parsing by Chunks” [A], In *Robert Berwick, Steven Abney and Carol Tenny (eds.) Principle-Based Parsing* [C], *Kluwer Academic Publishers*.

[19] L. Marquez, X. Carreras, K.C. Litkowski, and S. Stevenson. (2008) Semantic Role Labeling: An Introduction to the Special Issue. Computational Linguistics, 34(2): 145-159.

[20] Xavier Carreras1, Lluis Marquez, et. al. Learning and Inference for Clause Identification [A]. In Proc. of ECML'02 [C]. 2002.