中文词汇语义关系抽取评测大纲（修订版）

1．评测对象

本次评测的对象是中文词义语义关系（包括同义关系、上下位关系）抽取中的核心技术。

2．任务设置

本次评测包括 2 个子任务：同义词发现和下位词发现。

2.1 同义词发现

对给定词表中的每个词，本任务要求找出该词的同义词。同义词的来源不限定于给定词表，可以从其他资源（例如词典、互联网等）中获取。

同义词（同义异形词），指表达的意义相同或相近，但表达形式不同的词汇。其主要形式包括：

别名/俗称：包括书面语和口头语、学名和俗称、不同地区的称谓差异等。例如“计算机”和“电脑”互为同义词。“操作系统”（大陆）和“作业系统”（台湾）互为同义词。但仅为简体/繁体写法差别的，例如“计算机”和“計算機”，是同一个词汇，而不是同义词。

全称/简称：例如“中华人民共和国”与“中国”互为同义词。

异形词：指在普通话书面语中并存并用的同音、同义而书写形式不同的词语，例如“笔画”和“笔划”互为同义词。

外来语译名差异：外来语有时存在多种翻译形式，它们之间互为同义词。例如“奥巴马”和“欧巴马”互为同义词。

语义近似：指语义、语用上相近的词，例如“尊敬”和“敬重”互为同义词。

2.2 下位词发现

对给定词表中的每个词，本任务要求找出该词的下位词。下位词的来源不限定于给定词表，可以从其他资源（例如词典、互联网等）中获取。

下位词指其语义内涵包含在另一个词汇（称为上位词）内涵之中的词汇。即下位词是上位词的一个特殊实例。例如“水果”的下位词包括“苹果”、“梨”、“菠萝”等。“国家”的下位词包括“中国”、“美国”、“日本”等。“文本分类方法”的下位词包括“支撑向量机”、“贝叶斯分类”、“K 近邻”等。

本次评测中，下位词不包括采用一般限定语修饰给定词所构成的合成词（或词组）。例如“中国城市”不是“城市”的下位词。“红苹果”不是“苹果”的下位词。但专有名词不在此

列。例如“冠状病毒”仍认为是“病毒”的下位词。“红富士苹果”仍认为是“苹果”的下位词。

下位词不包括整体-部分关系。例如“车轮”不是“汽车”的下位词。“省”不是“国家”的下位词。

3. 评测方法

3.1 评测方式

本次评测为离线评测。参评单位自行处理数据，生成相应结果后提交。答案采用人工标注的方法确定。

3.2 评测步骤

1) 评测单位预先提供测试样例（包括答案）
2) 评测单位给出测试数据
3) 参评单位运行被测系统，得出测试结果
4) 参评单位提交测试结果
5) 评测单位标注答案，运行自动评测程序，统计评测结果

3.3 评测指标

评测采用三个指标：正确率（Precision），召回率（Recall）和 F 值（F-measure），分别计算其微平均和宏平均值。

3.3.1 微平均

微平均以每个语义关系为一个计算单元，具体计算公式如下：

正确率

表示发现的语义关系（同义或下位）中出现在标准结果中的比例，计算公式如下：

$$\text{正确率} = \frac{\text{发现的语义关系中出现在标准结果中的数量}}{\text{发现的语义关系总数}} \times 100\%$$

其中，词表中的每个词汇与发现的每个同义词（或下位词）为一条语义关系。发现的同义词之间的关系不计算在内。

召回率

表示标准结果中被正确发现的语义关系比例，计算公式如下：

$$\text{召回率} = \frac{\text{发现的语义关系中出现在标准结果中的数量}}{\text{标准结果中的语义关系总数}} \times 100\%$$

F 值

是正确率和召回率的调和平均数，计算公式如下

$$\text{F值} = \frac{2 \times \text{正确率} \times \text{召回率}}{\text{正确率} + \text{召回率}}$$

3.3.2 宏平均

宏平均以每个词为一个计算单元，每个词的评价指标计算公式如下：

$$\text{词}i\text{的正确率} = \frac{\text{发现的词}i\text{的语义关系在标准结果中的数量}}{\text{发现的词}i\text{的语义关系数量}} \times 100\%$$

$$\text{词}i\text{的召回率} = \frac{\text{发现的词}i\text{的语义关系在标准结果中的数量}}{\text{标准结果中词}i\text{的语义关系数量}} \times 100\%$$

$$\text{词}i\text{的F值} = \frac{2 \times \text{词}i\text{的正确率} \times \text{词}i\text{的召回率}}{\text{词}i\text{的正确率} + \text{词}i\text{的召回率}}$$

宏平均值计算公式如下：

$$\text{正确率} = \frac{1}{N}\sum_i \text{词}i\text{的正确率}$$

$$\text{召回率} = \frac{1}{N}\sum_i \text{词}i\text{的召回率}$$

$$\text{F值} = \frac{1}{N}\sum_i \text{词}i\text{的F值}$$

其中，N 为评测词汇总数。

3.4 数据集

同义词评测数据集包含 10000 个词汇。数据来源包括普通词典、百科词条、叙词表等多种资源。词汇的词性包括普通名词、专有名词、动词和形容词。

下位词评测数据集包括 10000 个词汇。数据来源包括普通词典、百科词条、叙词表等多种资源。词汇的词性包括普通名词和专有名词。

3.5 数据格式

所有数据文件均采用 utf-8 编码。

- 测试数据文件格式为：

测试语料 ::= {词汇\n}

词汇 ::= 字符{字符}

其中，{}表示重复项（可重复 0 次或多次）。即每行一个词汇。

- 需提交的同义词识别结果格式为：

结果 ::= {同义关系}

同义关系 ::= 原词汇{\t 同义词}\n

同义词 ::= 字符{字符}

即每行包含测试词表中的原词汇和它的所有同义词。词汇之间用制表符(‘\t’)分割。若词汇 t_1 和 t_2 都出现在测试词表之中，且互为同义词，则结果中应包含 *t1* 和 *t2* 为原词汇的两行结果，不得省略其中任意一行。

- 需提交的下位词识别结果格式为

结果 ::= {下位关系}

下位关系 ::= 原词汇{\t 下位词}\n

下位词 ::= 字符{字符}

即每行包含原词汇和它的所有下位词。词汇之间用制表符(‘\t’)分割。

3.6 评测要求

参评单位应当采用自动的方法，发现词汇的同义词和下位词。参评系统应当预先训练模型、调整好所有参数，运行过程中不得有人工干预。本次评测不限制使用各种语义资源。对于每个子任务，参评单位至多提交 2 组结果。

4. 评测日程

2012/1/1-2/29： 起草评测大纲，征求各方意见；

2012/3/1-3/31： 修订完善评测大纲，确定评测数据；

2012/4/1： 发布评测任务，接受评测报名；

2012/5/4： 发布评测样例数据集；

2012/5/31： 评测报名截止；

2012/5/1-6/30： 构建评测数据集，制定标准答案；

2012/7/1： 发布评测数据集；

2012/7/31： 参评单位提交运行结果；

2012/8/1-8/31： 组织专家评测小组进行结果评判，发布评测结果；

2012/9/1-9/24： 征集评测论文；

2012/9/25-9/30： 确定受邀报告；

2012/10/29-11/6： 宣读报告，交流经验和技术。

5. 如何注册

参加评测的单位需要在接受报名的时间内到如下评测官网进行并填写相关信息。

http://tcci.ccf.org.cn/conference/2012/

如果你有任何关于本次评测的问题请发邮件至：huangxiaojiang@pku.edu.cn

6. 本次评测的组织

- **主办单位**

  中国计算机学会中文信息技术专业委员会（CCF TCCI）

- **承办单位**

  北京大学，MSRA

- **协办单位**

  数字出版国家重点实验室

- **评测委员会（按照姓氏拼音排序）**

  李寿山、刘群、万小军、韦福如、吴云芳、徐睿峰