FudanDNN-NLP2.1中文使用手册

（郑骁庆 复旦大学）

本使用手册介绍复旦大学计算机学院机器人研究实验室所开发的基于深度学习的中文自然语言处理工具FudanDNN-NLP2.1，该工具目前可用于中文分词、命名识别、词性标注、句子分类、语义分析。深度学习方法的优点在于不需要预先根据任务进行特征的选择（特征工程），系统所需参数较少（节省内存开销），并且实际使用远远快于其它相似性能的系统。

1. 新增功能

2.1较2.0版本，本次主要增加以下功能：

(1) 增加了带转移矩阵的双向LSTM模型，能够用于各类序列标注和语义分析任务。
(2) 增加了带动态*k*-max池化的卷积神经网络的句子语义表示模型，可以根据其产生的句子级语义表示完成各种分类任务。
(3) 针对同时处理多个事件时，条件随机场CRF（Conditional Random Fields）模型处理速度慢的问题，对CRF模型的训练和解码算法进行优化，使其处理速度与单一事件相同。
(4) 基于条件随机场CRF（Conditional Random Fields）的语义分析模块。
(5) 增加时期类型的识别和支持自定义领域词汇。
(6) 用于中文分词、命名识别和词性标注的神经网络使用经过预训练的字向量进行重新训练，性能进一步提高。字向量采用最终研究成果的适用于中文的预训练方法。
(7) 修正了之前版本中的一些Bug。

2.0较1.0版本，本次主要增加以下功能：

(1) 支持根据用户所准备的样本进行模型的重新训练和调整训练（在已经大量样本训练的基础上根据新样本进行精调）。
(2) 对网络结构和学习处法进行优化，加速模型训练时间。
(3) 基于条件随机场CRF（Conditional Random Fields）的语义分析模块。
(4) 支持自定义词汇和文本规范化（俚语替换）功能。
(5) 完善各模块之间的整合，提升各种任务的准确性。

2. 部署步骤

使用工具进行项目开发的基本流程如下：

(1) 获取并安装Java JDK1.7或以上版本。
(2) 获取FudanDNN-NLPv2.1.zip压缩包，并解压。
(3) 将解压后的“FudanDNN-NLPv2.1”目录下的目录及文件全部拷贝到工程。
(4) 将“package”目录下的“FudanDNN-NLP2.1.jar”包导入工程。
(5) 根据本手册第3节各种功能的详细使用说明进行应用开发。

注意：使用本工具时，所有文件都应采用 UTF-8 编码，集成开发环境和编辑工具也应使用 UTF-8 编码，不然可能出现因编码不一致所导致的错误。

3. 使用说明

3.1. 中文分词

3.1.1. 使用方法

使用中文分词功能的样例代码见“cn.edu.fudan.flow”包下的“WordSegmentorStart.java”，该功能已经集成了自定义词汇、俚语替换和命名识别功能。

自定义词汇和俚语替换通过配置“source”目录中以下9个文件实现：

自定义词汇类别	文件名	备注
人名	person.utf8	每个自定义词条一行
组织机构名	organization.utf8	同上
成语	idiom.uft8	同上
食物	food.utf8	同上
著名景点	scene.utf8	同上
设备名称	device.utf8	同上
作品名称	title.uf8	同上，可包括著作、影视作品等
领域词汇	domain.utf8	同上，应用领域相关的专用名词
俚语	slang.utf8	分为两列，前一列为俚语、每二列为规范用语，两者以空格分开。一组一行，每一行最前面加“#”号，会忽略该行，删除后恢复。

命名识别模块会对以下类型进行识别：

类别	英文标签	备注
俚语	SLANG	来自slang.utf8文件所定义的词（自动替换，结果中不再显示）
邮箱地址	EMAL	
网址	URL	
日期	DATE	
百分比	PERCENT	
度量	MEASURE	细分为：LENGTH（长度）、WEIGHT（重量）、SQUARE（面积）、VOLUMN（体积）、TEMPERATURE（温度）。
时间	TIME	
时期	PERIOD	
货币	CURRENCY	
手机号码	CELLPHONE	
座机号码	LANDLINE	
领域词汇	DOMAIN	来自domain.uft8文件所定义的词
成语	IDIOM	来自idiom.uft8文件所定义的词
食物	FOOD	来自food.utf8文件所定义的词
熟知地点	SCENE	来自scene.utf8文件所定义的词
作品名称	TITLE	来自title.uf8文件所定义的词
设备名称	DEVICE	来自device.uft8文件所定义的词
组织机构名	ORGANIZATION	合并命名识别模块结果与来自organization.utf8文件所定义的词
地名	LOCATION	
人名	PERSON	合并命名识别模块结果与来自person.utf8文件所定义的词
外文字符	FOREIGN	
数字	DIGIT	
标点	PUNCUTATION	

注意：当某一词汇同属于两种类型时，按上表所示先后顺序的优先级确定。如：同时属于 CELLPHONE 和 DIGIT，则识别为 CELLPHONE 类型。

识别人名、组织机构名、地名的命名识别模块将先于中文分词模块运行，并将识别结果作用于中文分词模。

3.1.2. 配置文件

中文分词功能的配置文件见“conf”目录下的“WordSegmentor.properties”，包括以下参数：

参数名称	默认值	备注
inputNetworkSettingFile	model/WindowConvolutionNetwork_seg_d300_w5_f1.class	指向用于中文分词的网络参数文件。默认文件的网络参数已经经过大量语料训练，可以实际使用。使用自定义样本重新训练网络后，可以用结果文件进行替换。
isCharacterLevel	true	中文情况设置为true（一般不要修改）
isResultWithTag	false	中文分词设置为false（一般不要修改）
isSegmentation	true	中文分词设置为true（一般不要修改）
isStandard	false	指示是否要对输入语句进行规范化处理（包括全角转半角、繁体转简体、数字和字母统一表示形式）。由于中文分词之前的预处理模块已完成上述工作，所以此处设置为false（一般不要修改）。

中文分词之前会对输入句子进行必要的预处理（包括全角转半角、繁体转简体、数字和字母统一表示形式等），预处理的配置文件见“conf”目录下的“Preprocess.properties”，包括以下参数：

参数名称	默认值	备注
domainFile	source/domain.utf8	指向自定义“领域”词汇的文件路径
idiomFile	source/idiom.utf8	指向自定义“成语”词汇的文件路径
foodFile	source/food.utf8	指向自定义“食物名称”词汇的文件路径
sceneFile	source/scene.utf8	指向自定义“景点名称”词汇的文件路径
titleFile	source/title.utf8	指向自定义“作品名称” 词汇的文件路径
deviceFile	source/device.utf8	指向自定义“设备名称” 词汇的文件路径
slangFile	source/slang.utf8	指向自定义“俚语”词汇的文件路径
organizationFile	source/organization.utf8	指向自定义“组织机构名称”词汇的文件路径
personFile	source/person.utf8	指向自定义“人名”词汇的文件路径
nerRecogizerFile	conf/NerRecognizer.properties	指向中文命名识别模块配置文件的文件路径，配置文件相关参数说明详见3.2.2节。
isBIOES	true	指示模型是否采用BIOES标签规范，BIOES规范使用如下的规则：B开头标签表示词的开始字符；I开头表示词的中间字符；E开头表示词的结束字符；S开头表示单独成词的字符；O开头表示与任务无关的字符。如isBIOES设置成“false”，则使用BIO标签规范，即用B替代S，用I替代E。各模块训练样本所采用的标签规范需要与此参数设置一致。默认使用BIOES规范。
isDeleteAllBlank	true	指示是否删除输入句子中的空格。如果设置成“true”，则删除所有中间空格（两个英文单词之间的空格会用特殊字符“□”替换）。如果设置成“false”，所有空格会用特殊字符“□”替换。
isCaseSensitive	true	指示在识别自定义词汇时是否对大小写敏感，如果设置成“true”，输入句子中出现“ibm”，而在source/organization.utf8文件中包含“IBM”词条，则“ibm”不会被识别成公司名，如果设置成“false”，则会被识别成IBM公司的名称。

3.1.3. 重新或调整训练中文分词网络

重新或调整训练运行“cn.edu.fudan.dnn”包下的“WindowConvolutionNetworkStart.java”，

网络训练配置文件见“conf”目录下的“windowConvolutionNetwork.properties”，包括以下参数（注意：尽管以下有些参数的值在特定的情况下会被忽略，但在程序启动时，会被程序读取，所以都需要给出相应的值）：

参数名称	默认值	备注
corpusFile	dataset/segmentation_corpus.utf8	指定训练样本的文件路径。按照默认值所指向文件的格式准备训练样本，每一行包括一个字符与这个字符的标注，每一句之间用一行空格相隔。
tokenFile	conf/characters.utf8	指定中文字符集的文件路径。这个文件一般应包含出现在训练样本中的所有字符。应确保以下字符出现在字符集中：“╠”（句首字符）、“╣”（句尾字符）、“□”（未知字符）、英文字母和阿拉伯数字。“conf/characters.utf8”是该文件的一个范本。
labelFile	dataset/segmentation_labels.utf8	指定任务标注集合的文件路径。
isReadEmbedding	false	指定网络是否使用事先准备好的字或词向量进行训练。如果这个值设为“true”时，确保embeddingFile设置成适当的值。
embeddingFile	embedding/Word2Vector.class	指定网络启动时需要读取字或词向量的文件路径（即使用经过训练的字或词向量来进行初始化网络）。如果isReadEmbedding设置成为“false”，则embeddingFile的取值将被忽略。如何正确生成初始化网络字或词向量的“embedding.class”文件见3.1.4节。
isExternalFeature	false	对于字或词向量，是否使用外部特征值。如果这个参数设置为“true”，需要准备外部特征值文件（由externalFeatureFile参数指定储存外部特征值的文件路径），并且tokenFile文件会被忽略，字符集从externalFeatureFile指定的文件中读取。
externalFeatureFile	conf/ charactersWithExternalFeature.utf8	指定字或词向量外部特征值文件路径。默认值所指向的文件是该文件的一个范本。
isReadNetworkSetting	false	指定网络参数初值是否从某个经过训练网络的参数中读取。（使用之前训练过的网络参数初始化，继续进行训练过程，即调整训练）。如果这个参数设置为“true”，应确保inputNetworkSettingFile设置成适当的值。注意：当这个参数为“true”时，以下参数的值将被忽略：tokenFile、labelFile、isExternalFeature、externalFeatureFile、isReadEmbedding、embeddingFile、

		featureDimension、internalFeatureDimension、externalFeatureDimension、windowSize和isIgnoreAlphabetNumber。
inputNetworkSettingFile	model/windowConvolutionNetwork_d100_w3_f1.class	指定网络启动时需要读取网络参数的文件路径。如果isReadNetworkSetting设置成为“false”，则inputNetworkSettingFile的取值将被忽略，网络参数初值将随机产生。
outputNetworkSettingFile	model/windowConvolutionNetwork_d100_w3_f1.class	指定网络完成训练后，保存网络参数的文件路径。
echoFile	result/windowConvolutionNetworkEcho.utf8	指定记录训练过程信息的文件路径。
featureDimension	100	指定字或词向量的维度，该值应等于internalFeatureDimension和externalFeatureDimension取值之和，不然程序会报错。
internalFeatureDimension	100	指定字或词向量本身特征维度。
externalFeatureDimension	0	指定字或词向量外部特征维度。
windowSize	3	指定窗口的大小。
featureMap	1	指定卷积层feature map的数量，实验表时，取1时一般就能达到较好的性能。
learningRate	0.05d	指定学习步长。该值越大，学习速度越快。但是取一个较小的值有利于保持网络学习的稳定性。
regularizationRate	0.0001d	指定Regularization参数值。该参数用于防止过拟合。
errorLimit	0.001d	指定期望的误差水平。期望的准确率等于（1－errorLimit）。当网络的标注误差小于设定的值，网络停止训练过程，并且输出相应的网络参数。
learningTimes	100	指定最大的迭代次数，即扫描整个训练样本的次数。当迭代次数超过该值，网络停止训练（即使没有达到期望的误差水平）。
isIgnoreAlphabetNumber	true	指定是否忽略不同英文字母和阿拉伯数字的差异。如果这个参数设置为“true”，则所有英文字母的向量表示都相同，所有阿拉伯数字的向量表示也都相同。
isAverage	false	指定卷积层产生的多个feature map值之后取平均或求和。当设置为“false”时，进行相加操作，否则求平均值。当featureMap值设置成1时，isAverage取值对网络训练没有影响。

中文分词模块所使用标签及其说明如下表所示：

分词标签	标签说明
B	词的开始字符

I	词的中间字符
E	词的结束字符
S	单独成词的字符

3.1.4. 使用字或词向量初始化网络

网络参数的初始值对性能的影响比较大，并且网络参数数量最多的部分是字或词向量。实验结果表明使用经过预训练的字或词向量来初始化网络是提高性能的有效方法。字或词的向量预训练可以使用类似Google的Word2Vector[1]等工具。

中文一般使用和训练字向量，类似英文一般使用和训练词向量。训练好的字向量放置于“embedding”目录下，文件格式见目录下的“Word2Vector.utf8”文件，每一行为一个字符（中文）或词（英文）和其向量表示的各维度值，字或词与每一维度的值之间用空格隔开。运行“cn.edu.fudan.corpus”包下“LookupTableGeneratorStart.java”，会产生符合网络训练初始化要求的“Word2Vector.class”文件。运行“LookupTableGeneratorStart.java”需要配置“conf”目录下的“LookupTabelGenerator.properties”文件，该配置文件包括以下参数：

参数名称	默认值	备注
embeddingTextFile	embedding/Word2Vector.utf8	指定保存经过预训练产生字或词向量文件的路径，格式参见默认指向文件。
embeddingFile	embedding/Word2Vector.class	指定转换后满足网络初始化要求的字或词向量文件路径。该文件即为网络训练时embeddingFile参数所指向的文件。
dimension	50	字或词向量的维度，应与预训练所使用的字或词向量维度一致，否则程序会报错。
tokenFile	conf/characters.utf8	指定字符（中文）或词（英文）集合的文件路径。如果该集合含有预训练字或词中没有出现的字或词，则会通过随机初始化补全这些缺失的字或词向量，以确保网络训练不会报错。

3.2. 中文命名识别

3.2.1. 使用方法

使用中文命名识别的样例代码见“cn.edu.fudan.flow”包下的“NamedIdentityRecognizerStart.java”，该功能已经集成了自定义词汇和俚语替换功能（相关内容详见3.1.1节）。

中文命名识别模块所使用标签及其说明如下表所示：

命令识别标签	标签说明
PER	人名
ORG	组织机构名
LOC	地名
O	其它

模型采用与中文分文联合标注的方法，即同时识别词及其所属对象。通过采用中文分词标签和命名识别标签相连的方式产生适合于联合标注的标签集合，如：“B_PER”表示人名的开始字符，其它依此类推。

3.2.2. 配置文件

中文命名识别功能的配置文件见“conf”目录下的“NerRecognizer.properties”，包括以下参数：

参数名称	默认值	备注
inputNetworkSettingFile	model/	指向用于中文命名识别的网络参数文件。

[1] https://code.google.com/archive/p/word2vec/

	WindowConvolutionNetwork_ner_d300_w5_f1.class	默认文件的网络参数已经经过大量语料训练，可以实际使用。使用自定义样本重新训练网络后，可用结果文件进行替换。
isCharacterLevel	true	中文情况设置为true（一般不要修改）
isResultWithTag	true	命名识别设置为true（一般不要修改）
isSegmentation	false	命名识别设置为false（一般不要修改）
isStandard	false	指示是否要对输入语句进行规范化处理（包括全角转半角、繁体转简体、数字和字母统一表示形式）。由于中文命名识别之前的预处理模块已完成上述工作，所以此处设置为false（一般不要修改）。

3.2.3. 重新或调整训练中文命名识别网络

重新或调整训练运行“cn.edu.fudan.dnn”包下的“WindowConvolutionNetworkStart.java”，网络训练配置文件见“conf”目录下的“windowConvolutionNetwork.properties”，包括类似3.1.3节所介绍的参数。其中：“corpusFile”应指向为中文命名识别任务所准备的训练样本文件，参考格式见“dataset”目录下的“ner_corpus.utf8”文件，“labelFile”则应指向中文命名识别任务所使用标签集合文件，参考格式见“dataset”目录下的“ner_labels.utf8”文件。网络训练时使用预训练字或词向量初始化方法见3.1.4节所述。

3.3. 中文词性标注

3.3.1. 使用方法

使用中文词性标注的样例代码见“cn.edu.fudan.flow”包下的“PosTaggerStart.java”，该功能已经集成了自定义词汇、俚语替换和命名识别功能（相关内容详见3.1.1节）。

中文词性标注模块所使用标签及其说明如下表所示：

词性标签	词类说明
D	数词
IJ	语气词或叹词
JJ	形容词
C	连词
M	量词
NT	时间名词
NR	专有名词
FW	外文词
I	成语或习语
NN	一般名词
U	助词
ON	拟声词
PU	标点
AD	副词
LC	方位词
PN	代词
V	动词
P	介词
X	其它

模型采用与中文分文联合标注的方法，即同时识别词及其所属词性。通过采用中文分词标签和词性标签相连的方式产生适合于联合标注的标签集合，如：“B_D”表示数词的开始字符，其它依此类推。

3.3.2. 配置文件

中文词性标注功能的配置文件见“conf”目录下的“PosTagger.properties”，包括以下参数：

参数名称	默认值	备注
inputNetworkSettingFile	model/ WindowConvolutionNetwork _pos_d300_w5_f1.class	指向用于中文词性标注的网络参数文件。默认文件的网络参数已经经过大量语料训练，可以实际使用。使用自定义样本重新训练网络后，可用结果文件进行替换。
isCharacterLevel	true	中文情况设置为true（一般不要修改）
isResultWithTag	false	词性标注设置为false（一般不要修改）
isSegmentation	false	词性标注设置为false（一般不要修改）
isStandard	false	指示是否要对输入语句进行规范化处理（包括全角转半角、繁体转简体、数字和字母统一表示形式）。由于中文词性标注之前的预处理模块已完成上述工作，所以此处设置为false（一般不要修改）。

3.3.3. 重新或调整训练中文词性标注网络

重新或调整训练运行“cn.edu.fudan.dnn”包下的“WindowConvolutionNetworkStart.java”，网络训练配置文件见“conf”目录下的“windowConvolutionNetwork.properties”，包括类似3.1.3节所介绍的参数。其中：“corpusFile”应指向为中文词性标注任务所准备的训练样本文件，参考格式见“dataset”目录下“postagging_corpus.utf8”文件，“labelFile”则应指向中文词性标注任务所用标签集合文件，参考格式见“dataset”目录下的“postagging_labels.utf8”文件。网络训练时使用预训练字或词向量初始化方法见3.1.4节所述。

3.4. 句子语义表示模型

本工具采用带动态 k-max 池化的卷积神经网络模型来产生句子的语义表示，然后在此表示的基础上来完成分类等任务。

3.4.1. 使用方法

使用句子分类的样例代码见“cn.edu.fudan.sentence”包下的“ConvolutionalSentenceModelDecoderStart.java”程序。对于输入的句子，可以判断句子所属的类别。

3.4.2. 训练样本准备

重新训练句子（或短文）分类模型，首先需要准备正负样本，负样本是指应用不感兴趣的句子。正样本的格式见“sentence”目录下的“sample_sentence.utf8”文件，样本每一行包括句子类型标签和句子内容，标签与句子内容之间用空格隔开。负样本的格式见“sentence”目录下的“sample_negative.utf8”，负样本不需要标注类型标签（系统默认为“E_UNKNOWN”）。

运行“cn.edu.fudan.corpus”包下“SentencePrepareStart.java”程序，该程序会产生句子分类模型训练所需要的文件：包括正负样本的文件（顺序随机打乱）；字符表；标签文件。

3.4.3. 句子模型训练

运行“cn.edu.fudan.sentence”包下“ConvolutionalSentenceModelStart.java”程序，配置文件见“sentence”目录下的“sentence.properties”，包括以下参数：

参数名称	默认值	备注
corpusFile	sentence/sentence_corpus.utf8	指定训练样本的文件路径。该文件由样本准备程序自动生成，见3.4.2节。

tokenFile	sentence/token.utf8	指定训练样本中出现字符的字符表。该文件由样本准备程序自动生成，见3.4.2节。
labelFile	sentence/classificationLabel.utf8	指定任务标签集合的文件路径。该文件由样本准备程序自动生成，见3.4.2节。
isReadEmbedding	false	指定网络是否使用事先准备好的字或词向量进行训练。如果这个值设为“true”时，确保embeddingFile设置成适当的值。
embeddingFile	embedding/Word2Vector.class	指定网络启动时需要读取字或词向量的文件路径（即使用经过训练的字或词向量来进行初始化网络）。如果isReadEmbedding设置成为“false”，则embeddingFile的取值将被忽略。如何正确生成初始化网络字或词向量的“embedding.class”文件见3.1.4节。
isExternalFeature	false	对于字或词向量，是否使用外部特征值。如果这个参数设置为“true”，需要准备外部特征值文件（由externalFeatureFile参数指定储存外部特征值的文件路径），并且tokenFile文件会被忽略，字符集从externalFeatureFile指定的文件中读取。
externalFeatureFile	conf/charactersWithExternalFeature.utf8	指定字或词向量外部特征值文件路径。默认值所指向的文件是该文件的一个范本。
isReadNetworkSetting	false	指定网络参数初值是否从某个经过训练网络的参数中读取。（使用之前训练过的网络参数初始化，继续进行训练过程，即调整训练）。如果这个参数设置为“true”，应确保inputNetworkSettingFile设置成适当的值。注意：当这个参数为“true”时，以下参数的值将被忽略：tokenFile、labelFile、isExternalFeature、externalFeatureFile、isReadEmbedding、embeddingFile、featureDimension、internalFeatureDimension、externalFeatureDimension、windowSize、featureMap、numberOfLayer、numberOfTopK和isIgnoreAlphabetNumber。
inputNetworkSettingFile	model/sentenceModel.class	指定网络启动时需要读取网络参数的文件路径。如果isReadNetworkSetting设置成为“false”，则inputNetworkSettingFile的取值将被忽略，网络参数初值将随机产生。

outputNetworkSettingFile	model/sentenceModel.class	指定网络完成训练后，保存网络参数的文件路径。
echoFile	sentence/sentenceModelEcho.utf8	指定记录训练过程信息的文件路径。
featureDimension	60	指定字或词向量的维度，该值应等于internalFeatureDimension和externalFeatureDimension取值之和，不然程序会报错。 这个值最好是2的某个幂的值，如：32、64、128等。
internalFeatureDimension	60	指定字或词向量本身特征维度。
externalFeatureDimension	0	指定字或词向量外部特征维度。
numberOfLayer	2	网络层数，一般取2已经足够。
windowSize	5/3	指定各层网络的窗口大小，之间用“/”隔开。
featureMap	2/3	指定各层网络卷积层feature map的数量，之间用“/”隔开。
numberOfTopK	5	最后一层网络的k-max采样的k值大小。
learningRate	0.05	指定学习步长。该值越大，学习速度越快。但是取一个较小的值有利于保持网络学习的稳定性。
regularizationRate	0.0001	指定Regularization参数值。该参数用于防止过拟合。
errorLimit	0.001	指定期望的误差水平。期望的准确率等于（1 − errorLimit）。当网络的标注误差小于设定的值，网络停止训练过程，并且输出相应的网络参数。
learningTimes	1000	指定最大的迭代次数，即扫描整个训练样本的次数。当迭代次数超过该值，网络停止训练（即使没有达到期望的误差水平）。
isIgnoreAlphabetNumber	true	指定是否忽略不同英文字母和阿拉伯数字的差异。如果这个参数设置为“true”，则所有英文字母的向量表示都相同，所有阿拉伯数字的向量表示也都相同。

3.5. 基于CRF的中文语义分析

基于CRF的中文语义分析分为仅处理单事件和同时处理多事件两个版本，对多事件语义分析需要句子分类作为预处理（即过滤不感兴趣的句了，并且对可能描述某个事件的句子判断其描述事件的类型），句子分类的预处理采用3.4节所述模型。使用单事件语义分析也要确保输入的句子都是描标目标事件的语句。以下先介绍单事件语义分析的训练和使用过程，然后说明进行多事件任务的使用方法。

3.5.1. 使用方法

使用基于CRF模型的中文单事件语义分析的样例代码见“cn.edu.fudan.flow”包下的“CRFSemanticAnalyzerStart.java”程序。语义分析利用中文分词、命名识别和词性标注的结果，分析出输入句子的事件类型和关键的属性值对。

运行“CRFSemanticAnalyzerStart.java”涉及以下参数（在该类中直接设置）：

参数名称	默认值	备注
preprocessFile	conf/Preprocess.properties	指向预处理模块的配置文件路径，该配置文件参数详见3.1.2节

posTaggerFile	conf/PosTagger.properties	指向中文词性标注模块的配置文件路径，该配置文件参数详见3.3.2节
confFile	crf/crfDecoder.properties	指向中文语义分析模块的配置文件路径，该配置文件参数详见3.5.2节
eventKeywordFile	crf/eventKeywords.utf8	指向事件类型描述标签与中文解释之间对应关系的配置文件路径。配置文件中，每一行配置一个事件类型，事件类型描述标签与中文解释之间用空格隔开。
attributeKeywords	crf/attributeKeywords.utf8	指向关键属性描述标签与中文解释之间对应关系的配置文件路径。配置文件中，每一行配置一个关键属性，关键属性描述标签与中文解释之间用空格隔开。

3.5.2. 配置文件

中文语义分析功能的配置文件见“crf”目录下的“crfDecoder.propertiesproperties”，包括以下参数：

参数名称	默认值	备注
parameterFile	model/crf_demo.utf8	指向用于中文语义分析模型的参数文件。需要根据语义分析需求，使用自定义样本重新训练模型（详见3.5.3节），并将结果文件替换默认文件后才能实际使用。
templateFile	crf/template.utf8	指向条件随机场模型使用的特征模板文件的路径
labelFile	crf/semanticlabels_demo.utf8	指向中文语义分析任务所使用标签集合文件路径
is2gram	true	指示条件随机场模型是否使用前后标签转移特征，应对模型训练时的设置一致。
numberOfColumn	5	表示每一个词汇及其不同模块分析结果标签的总数，默认的取值为5（一般不要修改），包括：中文词汇（数字、网址、时间等用英文标签替换）、语义标签、词性标签、命名识别标签、中文原词形式。表示格式类似“crf”目录下“sample_demo.utf8”文件所示。

3.5.3. 训练基于CRF的中文语义分析模型

3.5.3.1. 准备训练样本

运行“cn.edu.fudan.corpus”包下“SemanticCorpusPrepareStart.java” 程序，该程序涉及以下参数（在该类中直接设置）：

参数名称	默认值	备注
preprocessFile	conf/Preprocess.properties	指向预处理模块的配置文件路径，该配置文件参数详见3.1.2节
posTaggerFile	conf/PosTagger.properties	指向中文词性标注模块的配置文件路径，该配置文件参数详见3.3.2节
sentenceFile	dataset/single_event_sentence.utf8或 dataset/multiple_event_sentence.utf8	指向存储将作为语义分析模型训练样本原始语句的文件路径。对于单事件，该文件包含描述事件的句子，每行一句。对于多事件，每行句子前附加事件类型标签（标签的集合应与句子分类模型使用的一致），事件标签和句子之前用空格隔开。

outputFile	dataset/single_event_sample.utf8或 dataset/multiple_event_sample.utf8	指向保存原始语句转化成训练样本格式的文件路径。格式为一个词汇一行，每行共5列，依次分别表示中文词本身（数字、网址、时间等用英文标签替换）、语义标签（此时统一为“O”，表示与任务无关内容，之后需要根据任务进行人工标注）、词性标签、命名识别标签、中文原词形式。每一句样本之间由一空行隔开。
isMultiEvent	true或false	多事件时为true，多事件时为false。
hasHeadingEventType	true或false	多事件时为true，多事件时为false。

对转化成训练样本格式的文件（outputFile所指向的文件）进行语义标注，即使用属性标签修改文件中第二列的语义标签。

3.5.3.2. 准备语义标签集合

将语义标注采用的所有标签汇总保存在“crf”目录下“semanticlabels_demo.utf8”文件中，以每一个语义标签一行的格式存储（包括“O”标签）。

3.5.3.3. 确定模型的特征模板

编辑“crf”目录下的“template.utf8”文件，该文件定义了依据哪些特征进行语义分析。具体内容如下所示（可以使用所提供的默认模板）：

```
# Unigram
U00:%x[-2,0]
U01:%x[-1,0]
U02:%x[0,0]
U03:%x[1,0]
U04:%x[2,0]
U05:%x[-2,0]/%x[-1,0]
U06:%x[-1,0]/%x[0,0]
U07:%x[-1,0]/%x[1,0]
U08:%x[0,0]/%x[1,0]
U09:%x[1,0]/%x[2,0]
U10:%x[0,0]/%x[0,2]
U11:%x[0,0]/%x[0,3]
U12:%x[0,0]/%x[-1,2]/%x[0,2]
U13:%x[0,0]/%x[0,2]/%x[1,2]

# Bigram
B00:%x[-2,0]
B01:%x[-1,0]
B02:%x[0,0]
B03:%x[1,0]
B04:%x[2,0]
B05:%x[-1,0]/%x[0,0]
B06:%x[-1,0]/%x[1,0]
B07:%x[0,0]/%x[1,0]
B08:%x[0,0]/%x[0,2]
B09:%x[0,0]/%x[0,3]
```

特征模板共有两类，分别以“U”和“B”开头。其中：“U”类特征模板所产生的特征仅考虑当前的语义标签，而“B”类特征模板所产生的特征考虑语义标签之间的转移关系。类似“[-1, 2]”表示利用哪些特征值来对当前的语义标签进行预测，其中“-1”（相对行数）表示利用前一个词的相关特征值，“2”表示（相对列数）哪一列的特征值，“[-1, 2]”则表示前一个词的词性标签。类似“[-1, 2]”可进行自由组合，例如：“%x[0,0]/%x[0,2]/%x[1,2]”，表示联合使用当前词汇、当前词汇词性和后一个词汇词性对当前的语义标签进行预测。

3.5.3.4. 模型训练

运行“cn.edu.fudan.crf”包下的“ConditionalRandomFieldStart.java”程序，其配置文件见“crf”目录下的“crf.properties”，包括以下参数：

参数名称	默认值	备注
corpusFile	crf/sample_demo.utf8	指向语义分析训练样本的文件路径。训练样本准备见3.5.3.1节
parameterFile	model/crf_demo.utf8	指定模型完成训练之后，保存模型参数的文件路径。
templateFile	crf/template.utf8	指向条件随机场模型使用的特征模板文件的路径
labelFile	crf/semanticlabels_demo.utf8	指向中文语义分析任务所使用标签集合文件路径。
numberOfColumn	5	表示每一个词汇及其不同模块分析结果标签的总数，默认的取值为5（一般不要修改），包括：中文词汇（数字、网址、时间等用英文标签替换）、语义标签、词性标签、命名识别标签、中文原词形式。表示格式类似“crf”目录下“sample.utf8”文件所示。
learningTimes	1000	指定最大的迭代次数，即扫描整个训练样本的次数。当迭代次数超过该值，模型停止训练（即使没有达到期望的误差水平）。
errorLimit	0.0001	指定期望的误差水平。期望的准确率等于（1－errorLimit）。当模型的标注误差小于设定的值，模型停止训练过程，并且输出相应的模型参数。
is2gram	true	指示条件随机场模型是否使用前后标签转移特征。

3.5.4. 基于CRF的多事件中文语义分析

3.5.4.1. 准备训练样本

见3.5.3.1节，对转化成训练样本格式的文件进行语义标注，与单事件不同是的，每一个样本前一行需要增加事件类型标签。样例文件如“crf/sample_multiple_demo.utf8”所示。

3.5.4.2. 准备语义标签文件

准备语义标签文件，其样例文件如“crf/semanticlabels_multiple_demo.utf8”所示，与单事件语义标签不同在于：不同事件的标签之间用一空行隔开，同一事件的语义标签以事件类型标签开始，然后列出所有该事件所有的属性标签。

3.5.4.3. 准备语义CRF的特征模板

参见3.5.3.3节说明。

3.5.4.4. 模型训练

运行“cn.edu.fudan.crf”包下的“ConditionalRandomFieldMultipleEventStart.java”程序，其配置文件见“crf”目录下的“crf_multiples.properties”，参数作用基本与单事件相同，其说明见3.5.3.4节。

注意：如果模型一般都会正常收敛，如发现不收敛，先检查语义标签文件是否正确完整，然后检查训练样本标注是否一致。

3.5.4.5. 训练描述事件的句子分类模型

见3.4节说明，注意。

3.5.4.5. 模型使用

使用基于CRF模型的中文单事件语义分析的样例代码见“cn.edu.fudan.flow”包下的“CRFSemanticAnalyzerMultipleEventStart.java”程序。

3.6. 基于LSTM的中文语义分析

LSTM（Long-Short Term Memory）模型能够综合长短程信息，特别适用类似语义分析等高层自然语言处理任务。带转移矩阵的双向LSTM优势在于：对每一个位置的标注，综合了全句完整信息和标签转移概率，利于获得全局的最优解。

采用LSTM模型进行语义分析时，需要为每一类事件训练相应的网络，每一个时刻网络的输入是词、词性、命名识别类型转化成相应向量的并合。

3.6.1. 训练样本准备

训练样本与使用 CRF 模型基本相同，只是需要为每一事件准备相应的训练样本和语义标签文件（类似 CRF 的单事件语义分析）。训练样本例示文件如“lstm/E_DATE_corpus.utf8”所示，相应的语义标签文件如“lstm/E_DATE_attribute.utf8”。

3.6.2. 训练LSTM网络

运行“cn.edu.fudan.lstm”包下的“BiLSTMwithTransitionStart.java”程序，其配置文件见“lstm”目录下的“bi_lstm.properties”。包括以下参数：

参数名称	默认值	备注
inputUnits	50	词汇、词性和命名识别类型转换成向量后的总维度数，即各embeddings维度之和。该值等于vocabularyDimension、posLabelDimension、preprocessLabelDimension三者之和
vocabularyDimension	30	词向量的维度
posLabelDimension	10	词性的维度
preprocessLabelDimension	10	命名识别类型的维度
outputUnits	6	事件属性语义标签的个数
numberOfBlock	100	LSTM网络中Block的数量，一般要多于事件属性语义标签的个数。
numberOfCell	1	每个Block的Cell数量，在同一个Block中的Cell共享输出、输出和遗忘三个门。
corpusFile	lstm/E_DATE_corpus.utf8	指向语义分析训练样本的文件路径。训练样本准备见3.6.1节。

outputNetworkSettingFile	model_lstm/E_DATE_lstm.class	指定模型完成训练之后，保存模型参数的文件路径。
semanticLabelFile	lstm/E_DATE_attribute.utf8	指向中文语义分析任务所使用属性标签集合文件路径。
vocabularyFile	lstm/vocabulary_lstm.utf8	在训练样本中出现的词汇集合，即为训练样本第一列出现词汇集合。不要忘记增加“□”符号。
posLabelFile	lstm/poslabels_lstm.utf8	在训练样本中出现的词性标签的集合，即为训练样本第三列的集合。样例文件所含标签已经齐全，无须修改。
preprocessLabelFile	lstm/preprocesslabels_lstm.utf8	在训练样本中出现的命名识别类型标签的集合，即为训练样本第四列集合。样例文件所含标签已经齐全，无须修改。
learningTimes	5000	指定最大的迭代次数，即扫描整个训练样本的次数。当迭代次数超过该值，模型停止训练（即使没有达到期望的误差水平）。
errorLimit	0.0001	指定期望的误差水平。期望的准确率等于（1－errorLimit）。当模型的标注误差小于设定的值，模型停止训练过程，并输出相应的模型参数。
learningRate	0.05	指定学习步长。该值越大，学习速度越快。但是取一个较小的值有利于保持网络学习的稳定性。
regularizationRate	0.0001	指定Regularization参数值。该参数用于防止过拟合。

注意：如果模型一般都会正常收敛，如果发现不收敛，先检查训练样本标注是否一致，然后尝试调整网络的超参数。LSTM训练一般较CRF需要较多的迭代次数。

3.6.3. 使用基于LSTM的单事件语义分析模型

运行“cn.edu.fudan.flow”包下的“LSTMSemanticAnalyzerStart.java”程序。

3.6.4. 基于LSTM的多事件语义分析训练过程

首先根据上述说明恰当设置各配置文件，然后训练描述事件的句子分类模型（见3.4节），最后为每一类事件训练一个 LSTM 网络。

3.6.5. 使用基于LSTM的单事件语义分析模型

运行“cn.edu.fudan.flow”包下的“LSTMSemanticAnalyzerMultipleEventStart.java”程序。其中需要增加一个“lstm/bi_lstm_multiple.properties”配置文件，该文件其实上将事件类型标签对应到相应的经训练的 LSTM 网络参数文件。

4. 其它信息

工具使用过程中，如有任何问题或建议，请通过邮件 zhengxq@fudan.edu.cn（郑骁庆）联系我们。

参考文献

[1] Xiaoqing Zheng, Jiangtao Feng, Mengxiao Lin, Wenqiang Zhang. Context-specific and multi-prototype character representations. *In Proc. The Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI'16)*, 2016.

[2] Xiaoqing Zheng, Hanyang Chen, Tianyu Xu. Deep learning for Chinese word segmentation and POS tagging. *In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'2013)*, Seattle, Washington, USA, 18-21 October, 2013, pp. 647–657.

[3] Xiaoqing Zheng, Haoyuan Peng, Yi Chen, Pengjing Zhang, Wenqiang Zhang. Character-based parsing with convolutional neural network. *In Proc. The Twenty-Fourth International Joint Conference on Artificial Intelligence (IJCAI'15)*, 2015.

[4] Nal Kalchbrenner, Edward Grefenstette, Phil Blunsom. A convolutional neural network for modelling sentences. *In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL'14)*.

[5] Michael Collins. 2002. Discriminative training methods for hidden Markov models: Theory and experiments with perceptron algorithms. *In Proceedings of the International Conference on Empirical Methods in Natural Language Processing (EMNLP'02)*.

[6] Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12: 2493– 2537.

[7] John Lafferty, Andrew McCallum, and Fernando Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. *In Proceedings of the International Conference on Machine learning (ICML'01)*.

[8] Hwee T. Ng and Jin K. Lou. 2004. Chinese part-of- speech tagging: one-at-a time or all-at-once? word-based or character-based? *In Proceedings of the Inter- national Conference on Empirical Methods in Natural Language Processing (EMNLP'04)*.

[9] Nianwen Xue. 2003. Chinese word segmentation as character tagging. *Computational Linguistics and Chinese Language Processing*, 8(1): 29–48.

[10] Yue Zhang and Stephen Clark. 2008. Joint word segmentation and pos tagging using a single perceptron. *In Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics (ACL'08)*.