基于特征融合和 PCNN 的实体关系抽取模型研究

学号：19124418 姓名：纪斌斌

1 研究问题

1.1 研究背景

随着网络的普及提速以及大量智能终端的出现，海量的数据信息每时每刻都在涌现，其中绝大多数的信息会以自然语言的形式存在，但这些数据不能自动呈现有价值的信息。因此，如何利用这些海量数据成为主要研究内容。本文关注于实体关系抽取，主要目的是发现各知识要素之间存在的关系，他提供了一种细粒度的语义信息，被用于自动问答系统、机器翻译、知识推理、文摘等众多方面。例如在语言标志任务中，实体关系识别可以辅助关联语义网络的知识单元建立。其自身的技术发展将大大带动其它自然语言处理技术的进步，因此具有很大的研究和应用价价值。

1.2 研究问题特点及难点

实体关系抽取需要从一个同时带有两个实体的句子中，识别出两个实体的关系，该类问题通常可以被转换为关系分类问题。相较于传统文本情感分类的问题，该问题需要准确识别句子和两个实体的关系，问题难点主要集中于[1]：

- 文本的向量化表达。在该问题中，可同时考虑序列特征和结构特征，不仅利用词向量模型进行训练，还可添加词性标注、上位词标注以及句法依存等；
- 理解各实体与句子存在的关系。要解决该问题不仅需要通过模型提取全局信息和局部信息，实体与句子的关系提取，两实体前、中、后部分的语句信息也十分重要；
- 分类器以及损失函数选择。

1.3 研究内容

本研究关注于句子级别的实体关系抽取，对比分析了融合实体特征抽取和注意力机制的 BiGRU 模型、基于上下文感知融合分段池化的 BiGRU 模型、基于特征融合的 PCNN 模型三种模型的分类能力。结果显示基于上下文感知融合分段池化的 BiGRU 模型效果最佳。该模型将句子中的每个单词与两个实体以及位置信息进行拼接，并使用分段池化抽取两个实体前、中、后部分的语义信息，利用 softmax 分类器进行分类。最后，使用精确率、召回率、F1 值三个指标，同时使用宏观的加权平均两种评价方式对各个模型分类结果进行评价。

2 研究数据

2.1 数据基本信息

本文选取的数据集包含 101717 个带有实体信息和关系类型的实例，其中 8000 个训练样本，2717 个测试样本。在每个样本中包含 2 个有关系和实体，且这 2 个实体仅属于 1 中关系类型，该数据集中共有 9 中带有方向的关系和 1 种不带方向的 other 类型关系。表 1 展示了该数据集的数据样例。

表 1 数据样例

	e1	e2	Sentence	relation
1	composer	oblivion	Their composer has sunk into oblivion	Other
2	People	downtown	People have been moving back into downtown	Entity-Destination(e1,e2)

2.2 数据预处理

本文的数据预处理主要包括:

- 大小写统一、缩写更正、符号删除以及分词操作；
- 实体位置确定。该数据集标明了实体单词，而并未准确地在句子中标识实体位置。因此，为确定实体位置主要使用了以下几点位置标注原则：
  1) e1 的位置严格前于 e2；
  2) 若 e1、e2 中出现词组，则标注时各个单词应该是不被分割的连续单元；
  3) 若多个 e1、e2 同时满足上述条件，则标号大对大小对小组成几个实体对，最后选择两实体间距离较大的一对。
- 利用“glove.6B.100d”的字典进行词语编码；
- 使用 LabelEncoder 对类型标签进行编码。

最后，经过处理后的数据如表 2 所示：

表 2 数据样例

	e1	e2	Sentence	relation
1	[50]	[52]	[15, 10, 2929, 206, 14, 179, 599, 2449, 114…	13
2	[15]	[18,19]	[0, 436, 1061, 10797, 670, 16, 3088, 18361...	2

3 模型设计

3.1 CNN 模型基本原理

CNN 的本质在于构建多个能够提取数据特征的滤波器，通过对输入数据进行逐层卷积和池化操作来提取数据之间隐藏的拓扑结构特征。随着层数的不断增加，提取的特征越来越抽象，最后将这些抽象特征通过全连接层汇合，并通过 softmax 或 sigmoid 激活函数解决分类问题和回归问题。CNN 的特点之一在于可提取输入数据的局部特征，并逐层组合抽象生成高层特征，有效实现隐藏特征提取。

3.2 融合位置特征的 PCNN 模型建立

该模型基于 CNN 进行实体关系抽取，主要关注于文本上下文特征、实体与单词产生的交互特征以及单词与两实体的位置特征。具体模型结构如图 1 所示，主要流程如下：

- 通过查找预训练的词向量表，生成每个句子的词向量矩阵以及 2 个实体词向量，将两个实体词向量同时与句子中的每个词语进行拼接[3]；
- 加入每个单词的位置特征向量，对于一个句子的词序列$x = [x_1, x_2, x_3, \dots x_n]$中的每个词 x_i，其距离 2 个实体的相对距离为 $i - i_1$和 $i - i_2$，i 表示该词在

当前句子中的索引，i_1和 i_2分别表示 2 个实体的索引，组成最终的词向量矩阵[2]；

- 通过 CNN 卷积运算得到一系列特征，根据各个单词在句子中的相对位置，把句子级别的特征向量分为三部分[4][5]：
  1) before：第一个实体之前的所有上下文，包括第一个实体；
  2) middle：两个实体之间的上下文信息，包括两个实体；
  3) after：第二个实体之后的上下文信息，包括第二个实体。
- 在池化层的作用下获得各个部分固定长度的特征向量，组合成最终特征向量。
- 最后通过全连接层进入到 softmax 分类器中进行分类。

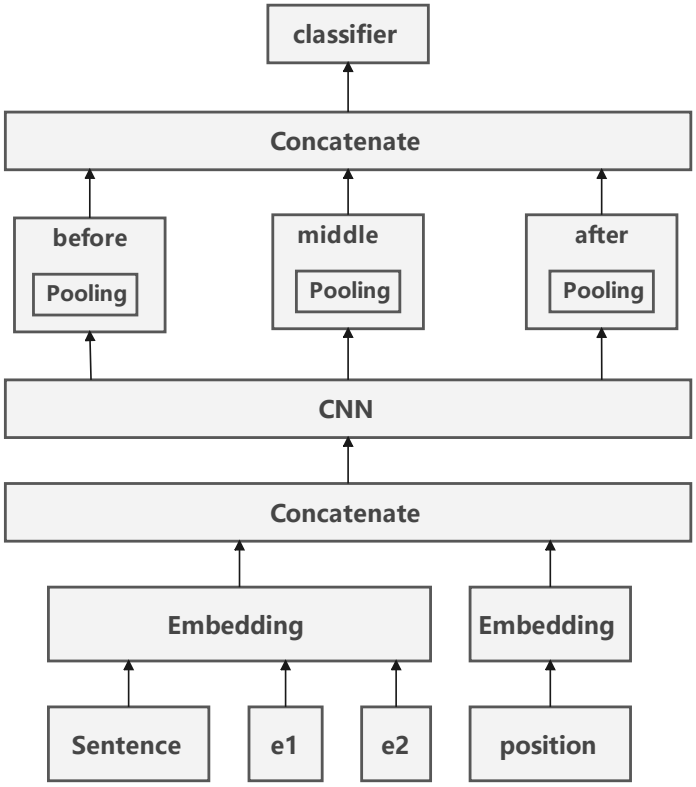

图 1 基于特征融合的 PCNN

4 实验设计

4.1 实验对比模型

4.1.1 融合实体特征抽取和注意力机制的 BiGRU 模型（BiGUR-1）

该模型重点关注于两个实体与整个句子的交互关系特征提取。主要流程为：

- 通过查找预训练的词向量表，生成每个句子的词向量矩阵；
- 将词向量矩阵输入 BiGRU 模型中提取句子级别和词语级别的特征；
- 将两个实体词向量向量分别与 BiGRU 的每个隐藏层输出进行拼接；
- 使用 attention 得出各实体与句子交互产生的特征向量，其中，attention 的具体计算方式如下所示，其中H表示 BiGRU 每个隐藏层输出与实体拼接后的向量表达，h^*为最终的表达[6]：

$$M = \tanh(H) \tag{1}$$
$$\alpha = \text{softmax}(W^T H) \tag{2}$$
$$\gamma = H\alpha^T \tag{3}$$
$$h^* = \tanh(\gamma) \tag{4}$$

- 将两个实体交互特征向量与 BiGRU 最后一层的输出进行拼接得到最终的特征向量，最后方式 softmax 分类器中完成分类。

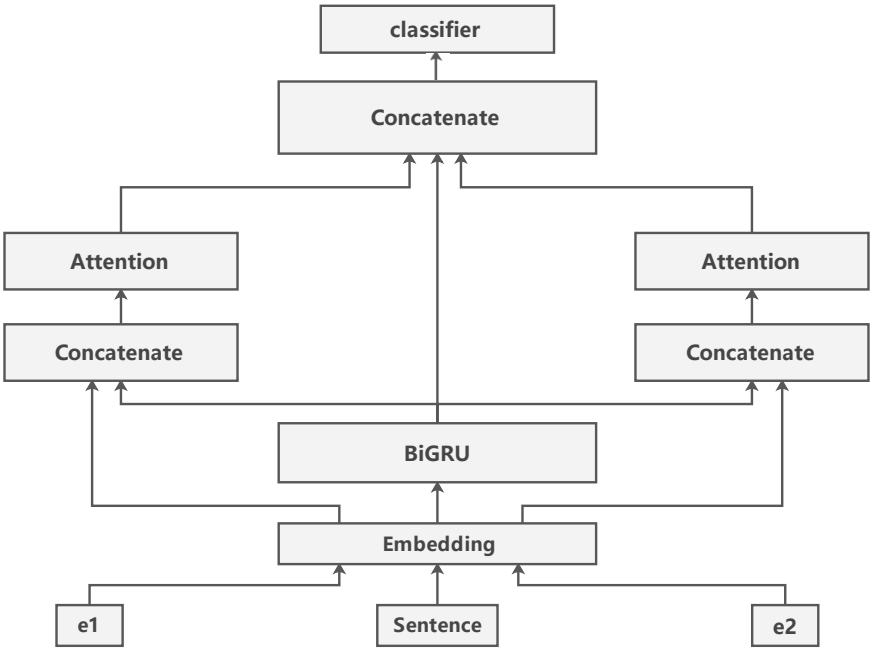

图 2 融合实体特征抽取和注意力机制的 BiGRU 模型（BiGRU-1）

4.1.2 基于上下文感知融合分段池化的 BiGRU 模型构建（BiGRU-2）

本文提出的基于上下文感知的 BiGRU 模型，主要思想为同时关注句子的上下文特征、全局特征以及两实体的特征，模型的整个流程如下：

- 数据预训练，抽取句子中词法级别的特征，将词语向量化表达；
- 将词的各个特征向量级联，形成最终输入向量，放入 BiGUR 模型，生成最终句子级别的特征向量序列；
- 根据各个单词在句子中的相对位置，把句子级别的特征向量分为前、中、后三部分[3]；
- 分段池化获得各个部分固定长度的特征向量，与 BiGUR 最后一个时间步的隐藏层状态以及两实体特征向量进行向量联级，得到最终的特征向量；
- 将特征向量放入 softmax 分类起于此实体对具体的关系类型。

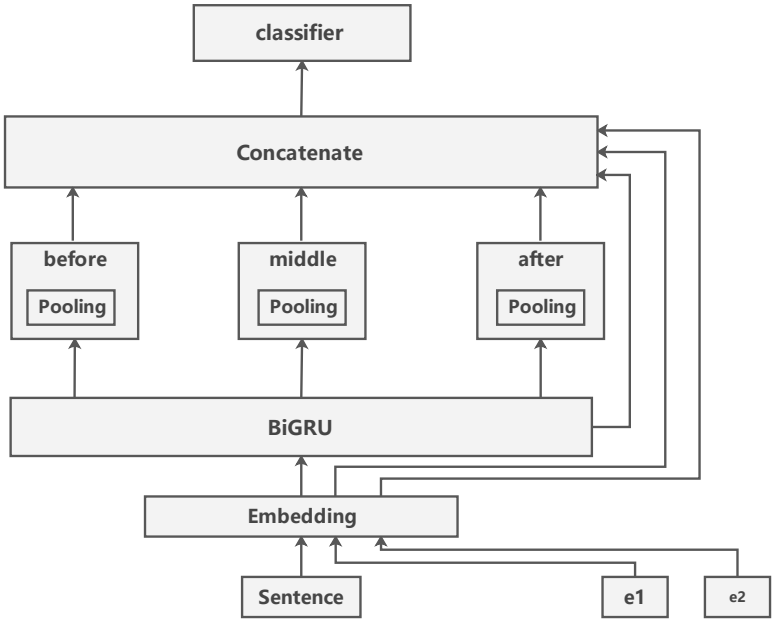

图 3 基于上下文感知融合分段池化的 BiGRU 模型（BiGRU-2）

4.2 模型评价指标

文研究问题为多分类问题，指标选用精确率、召回率、F1 值，同时使用宏观的加权平均两种评价方式。评价时分别将各类别的评价转换为二分类问题评价，计算相关指标，最后用求算术平均平均或加权平均的方式得出最终评价结果。

4.2.1 精确率

精确率在该问题中，实际意义为识别为某种实体关系的样本中真正属于该关系的样本的比例。具体计算过程如下。

$$Precision_i = \frac{TP_i}{TP_i + FP_i} \tag{5}$$

4.2.2 召回率

召回率在该问题中，实际意义为某种实体关系的样本中被预测为该关系样本的比例。具体计算过程如下。

$$Recall_i = \frac{TP_i}{TP_i + FN_i} \tag{7}$$

4.2.3 F1 值

因为 Precision 和 Recall 是一对相互矛盾的量，当 Precision 高时，Recall 往往相对较低，当 Recall 高时，Precision 往往相对较低，所以为了更好的评价分类器的性能，一般使用 F1-Score 作为评价标准来衡量分类器的综合性能。具体计算过程如下。

$$\text{Micro F1 score} = \frac{2 * Micro\ Precision * Micro\ Recall}{Micro\ Precision + Micro\ Recall} \tag{9}$$

4.3 实验结果

PCNN 模型使用 Adam 优化器，交叉熵损失函数，参数设置如表 3 所示。

表 3 模型参数设置

	参数名称	参数值
1	batch_size	64
2	word_embedding_dim	100
3	position_embedding_dim	25
4	kernel size	3
5	filter_num	100
6	epoch	30
7	learning_rate	0.001
8	drouout	0.5
9	L2	1e-5

PCNN 训练过程中，训练集和测试集的损失函数曲线和准确率曲线如图 4 图 5 所示。模型在训练集损失值收敛于 0.5 附近，测试集损失值有波动且损失值较高。但从准确率的角度看，测试集准确率始终高于训练集准确率，并且最终测试集准确率最终达到 0.78 上下，故认为该模型能有效广泛提取文本及实体特征关系，且不会出现过拟合现象。

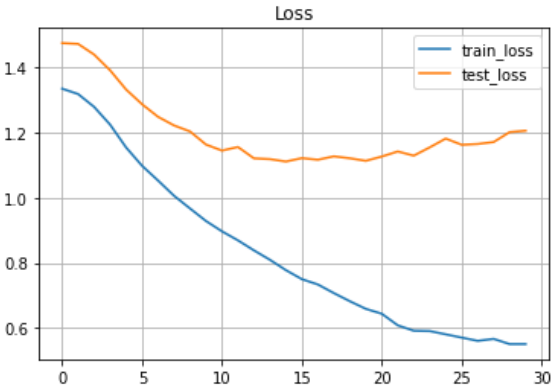

图 4 PCNN 损失函数曲线

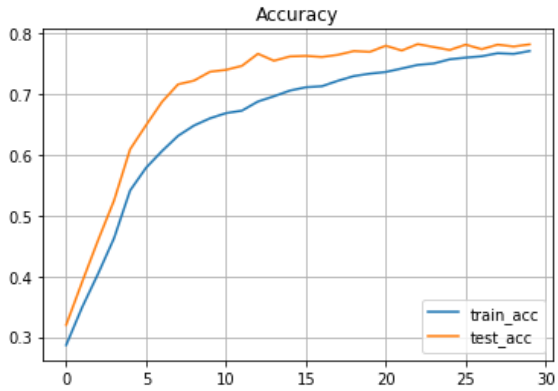

图 5 PCNN 准确率曲线

融合实体特征抽取和注意力机制的 BiGRU 模型（BiGRU-1）的损失函数曲线和准确率曲线如图 6 图 7 所示。模型在训练集损失值收敛于 0.25 附近，测试集损失值有波动且损失值较高。从准确率的角度看，训练集准确率接近于 1，测试集准确率最终达到 0.74 上下。认为该模型也能有效提取文本及实体关系，但存在过拟合的现象。

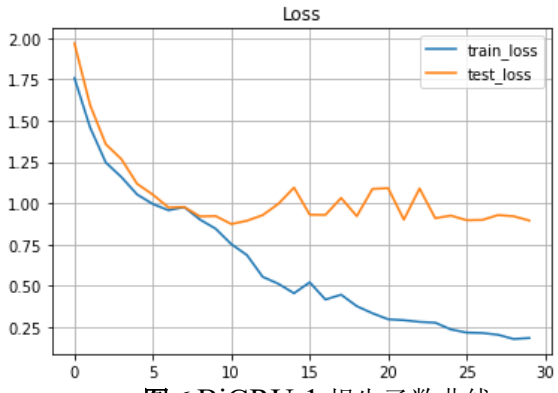

图 6 BiGRU-1 损失函数曲线

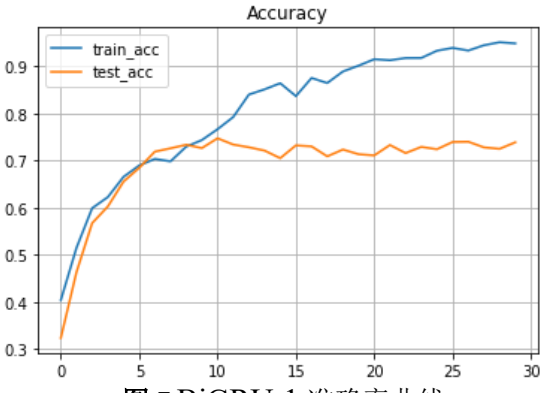

图 7 BiGRU-1 准确率曲线

基于上下文感知融合分段池化的 BiGRU 模型（BiGRU-2）的损失函数曲线和准确率曲线如图 8 图 9 所示。模型在训练集损失值收敛于 0.1 附近，测试集损失值有波动且损失值较高。从准确率的角度看，训练集准确率接近于 1，测试集准确率最终达到 0.7 上下。认为该模型在有效提取部分文本及实体关系是存在局限，且模型存在较为严重的过拟合的现象。

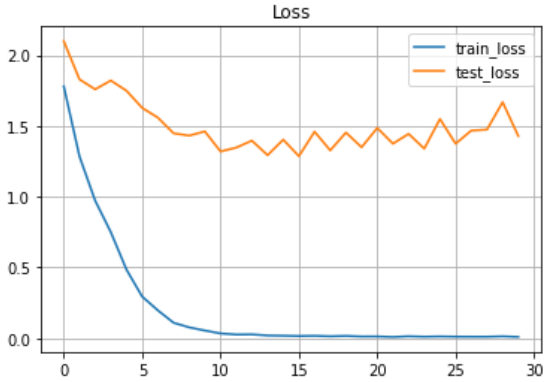

图 8 BiGRU-2 损失函数曲线

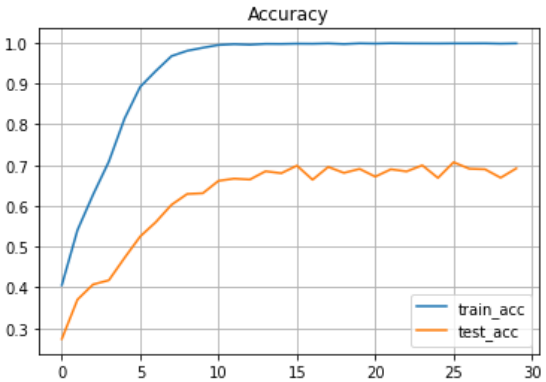

图 9 BiGRU-2 准确率曲线

4.4 模型对比

表 4 为各模型在测试集上的分类效果。从测试集的分类结果看，PCNN 的精准率、召回率以及 F1 值都最高，其次是基于上下文感知融合分段池化的 BiGRU 模型（BiGRU-2），最后是融合实体特征抽取和注意力机制的 BiGRU 模型（BiGRU-1）。从实验结果上看，加权平均的评价指标计算方式均高于算数平均，各类样本的数量偏差对模型影响较大。

表 4 模型对比

	Precision		Recall		F1-score	
	macro	weighted	macro	weighted	macro	weighted
BiGRU-1	0.5731	0.6995	0.4609	0.5745	0.5057	0.6266
BiGRU-2	0.6448	0.7505	0.5773	0.7300	0.5990	0.7361
PCNN	**0.6975**	**0.7729**	**0.6783**	**0.7729**	**0.6856**	**0.7715**

分析上表结果以及实验过程，可得出以下几点结论：

- 分段池化的有效性。PCNN 和 BiGRU-2 能相对较好的解决问题，二者都进行了分段池化的操作，提取了句子相对于两个实体前、中、后特征，实验证明该提取方式能有效理解实体和句子的关系。
- 融合位置信息的有效性。在 PCNN 的实验中，设计位置信息一方面是考虑迟到 CNN 模型在处理连续的时序数据时对长距离先后顺序的掌控能力较弱；另一方面也考虑位置信息在训练时可以提示该词语与实体的距离甚至相关性和重要度。但需要再优化的是位置信息的向量化表达。
- 实体信息与句子交互粒度越小，特征提取越有效。BiGRU-2 在加强实体与句子信息的交互中，使用较大的特征粒度，其主要是将实体与提取到的句子不同位置特征进行拼接运算；而 PCNN 使用词语级别的特征粒度，运算时直接在输入层将两个实体与各个单词的特征向量融合，结果显示特征融合粒度越小交互信息提取越有利。

- 越早将实体与句子进行交互效果可能越好。以上三个模型都有针对实体个句子的交互设计，但区别在于交互的时间及方式不同，从结果上看 PCNN 从网络输入开始便进行融合的方式能产生较好的效果。
- 文本的向量化表达在较为重要。该实验尝试使用 50 维、100 维、300 维的预训练词向量，结果表明同一模型使用 300 维词向量效果比 50 维能高出 1%左右，但由于电脑算力有限本实验选取 100 维词向量进行。在很多论文中，文本向量化表达，通常还会加上上位词语、词性标注、句法依存等特征。

5 未来改进方向

针对该实验未来改进方向主要集中于以下几点：

- 再次尝试 attention 在模型中的不同使用，在本次实验中，每个模型都尝试使用过不同的 attention 机制，有基于实体的单词相似性的，也有基于实体和特征向量交互的。但是实验中真正有效果的 attention 很少，部分 attention 加入后甚至导致运行时间变长、效果变差。需要重点考虑 attention 的交互设计原理及对象是否合理，以及使用 attention 的位置所产生的影响，即是在 embedding 后直接计算、网络输出结果后进行计算等。
- 丰富文本的向量化表达，可以参照论文加入上位词标注、词性标注、句法依存等信息。在实验中发现，就算搭建了和论文一致的模型，效果仍然不及论文，其中一个最大的一个问题就出现在文本向量化表达上。
- 关键词识别。本次实验原计划再进行完实体特征融合的交互设计后，便进行关键词识别的实验，但时间有限尚未完成。关键词识别在该问题中目前找到一些关键词识别方式，立即可执行的包括 TF-IDF 等，关键词识别后与句子的融合工作也值得研究。

6 参考文献

[1] 张少伟,王鑫,陈子睿,王林,徐大为,贾勇哲.有监督实体关系联合抽取方法研究综述[J].计算机科学与探索,2022,16(04):713-733.

[2] 王林玉,王莉,郑婷一.基于卷积神经网络和关键词策略的实体关系抽取方法[J].模式识别与人工智能,2017,30(05):465-472.

[3] Wang Y. et al., "Attention-based LSTM for aspect-level sentiment classification"[J], *Proc. Conf. Empirical Methods Natural Lang. Process*, 2016, vol.11, pp. 606-615.

[4] 任远方. 基于上下文感知的神经语义关系分类的研究[D].武汉大学,2017.2016, 49(1): 52–62.

[5] Santo C. et al. “Classifying Relations by Ranking with Convolutional Neural Networks”[J]. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing,* 2015, vol.1, pp626-624.

[6] Zhou P. et al., “Attention-Based Bidirectional Long Short-Term Memory Networks for Relation Classification”[J], *In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics,* 2016, vol(2), pp.207-211.