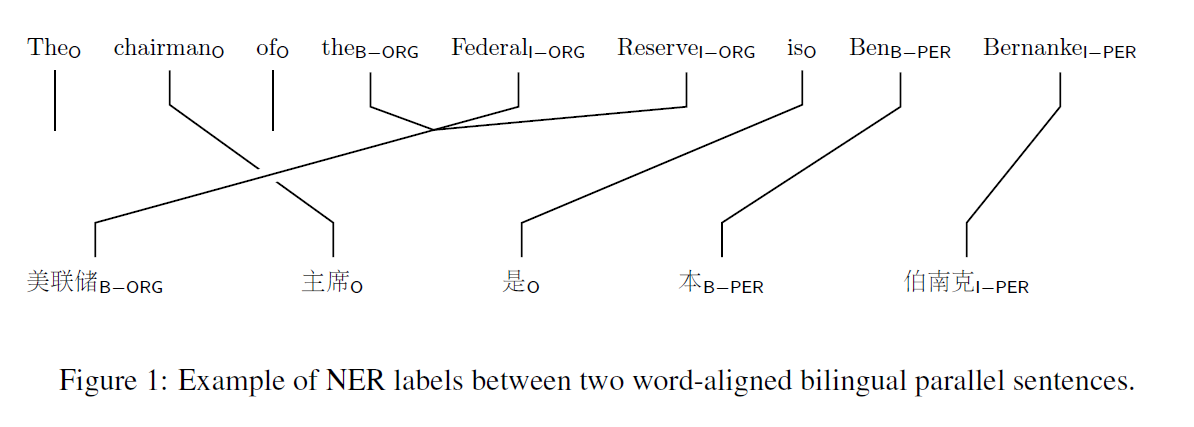
All Your Need is Attention for Bilingual Ner Mining

Hao Yang (Required)

2012 Labs，Huawei Co. Ltd  
fifayh@gmail.com



Abstract

Named entity missing and recognition errors are top issues in multilingual generation scenarios such as machine translation, accounting for 5.46% and 4.38% respectively (Hassan, et al. 2019). Bilingual constraints (the category and the number of detected entities must be equal for bilingual pair) can be defined as SoftTag (Che, et al. 2013) and SoftAlign(Bahdanau, et al. 2016). We propose Multilingual Bert with Tag and Align attention Net (META Net), by adding SoftTag and SoftAlign attention layers to learn the bilingual constraints. Experiments are conducted on English-Chinese and English-German with OntoNotes 5.0, ConLL2003 and WMT18 aligned corpus. All NER results are better than SOTA monolingual model, where Person Entity F1 in Chinese is significantly improved with 3.3%. We also have following findings from case studies. (1) Strong features such as capitalized Person Name in English and Location with common suffixes in Chinese can be learned by bilingual NER transfer. (2) Because sentence-level pairs are derived from document-level or paragraph-level pairs, which often causes ambiguity and omission and finally becomes noise for bilingual text.

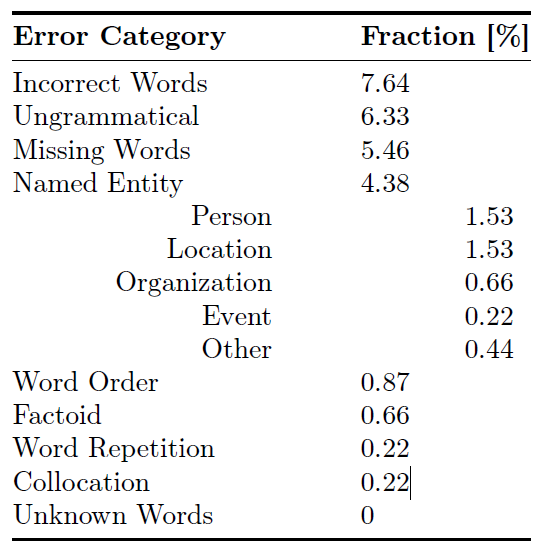
Deleting this Example of a Section Heading will Remove the Copyright Line*[[1]](#footnote-1)*

There are invisible characters after the word Line in the heading above. If you delete them, the copyright line will disappear. The AAAI copyright notice is required. This is example text. It is 10 point Times New Roman. The first sentence after the heading begins without a paragraph indent. Use of the text style (without alteration) is required.

This is example text. The second paragraph is indented 10 points, with no extra space between lines. The text is 10 point Times New Roman.

Introduction

From Keynote Report (Ming et al. 2019) in ACL 2019 “Achieving Human Parity on Automatic Chinese to English News Translation”, a detailed manual analysis is give that “Entity Missing” and “Named Entity Recognition (NER) Error” are top problems for multilingual generation scenarios and machine translation, in figure 1.



For bilingual NER in Statistical Machine Translation research (SMT), (Che et al. 2013) provides a soft-tag /soft-align architecture based on two constraints, monolingual constraints and bilingual constraints. As Figure 2, monolingual constraint is “inside entity(I-X) must follow begin entity(B-X) or inside entity(I-X)”, such as “… is [Ben]B-PER [Beranke]I-PER”, which can’t be “is [Ben]O [Beranke]”. Bilingual constraint is “align pair should be entity of same category”, which means pair sentences of align corpus should have same entity category and entity number. As “…is [Ben]B-PER [Beranke]I-PER” is person in English , “…是[本]B-PER[伯南克]-PER” should be person in Chinese. While because of one-hot presentation, large number manual CRF feature, and Berkeley Align, soft-tag and soft-align are not soft enough, and clustering dictionary is needed to enhance performance, which is not much stable.

Soft-Align is much softer in (Bahdanau et al) with an self-attention mechanism, and grows to be stacked as” Attention is all you need (Transformer)”([A Vaswani](https://scholar.google.com/citations?user=6rUjwXUAAAAJ&hl=en&oi=sra) et al). Unlike the traditional statistical machine translation, the neural machine translation aims at building a single neural network that can be jointly tuned to maximize the translation performance. As the use of a fixed-length vector is a bottleneck in improving the performance of NMT encoder–decoder architecture, and propose to attention weights by allowing a model to automatically (soft-)search for parts of a source sentence that are relevant to predicting a target word, without having to form these parts as a hard segment explicitly. This new approach achieved state-of-the-art (SOTA) BLEU score for almost all language pairs.

With the rapid development of deep learning, some neural network based methods for NER have been proposed and made a great success. (Collobert et al) early used CNN +CRF networks and (Schmidhuber et al) combined bidirectional LSTM(Bi-LSTM) with CRF. Former work provides two network structures, window method as CNN layer and sequence tagging as CRF layer. The window method (CNN) used the context window of the current predicted word as the input of the traditional linear network. CRF layer, as a un-direction graph model, predicate tagging with sequence feature as edge and current token feature as node. And objective functions was the wordlevel log-likelihood, which used softmax to predict the label probability. The latter work used Bi-LSTM as feature engineering which can learning bi-directional sequence features much better, and also used CRF to predict the label sequences.

Pretrained model such as BERT ([Pre-training of deep bidirectional transformers for language understanding](https://arxiv.org/abs/1810.04805)) and Multilingual BERT (M-BERT), released by Devlin et al. (2019) as a single language model pre-trained from monolingual corpora in 104 languages, is surprisingly good not only at many GLUE NLP tasks but also at zero-shot cross-lingual model transfer, in which task-specific annotations in one language are used to fine-tune the model for evaluation in another language. Experiments show that even to languages in different scripts, that transfer works best between typologically similar languages, that monolingual corpora can train models for code-switching, and that the model can find translation pairs. M-BERT does create multilingual representations, but that these representations exhibit systematic deficiencies affecting certain language pairs.

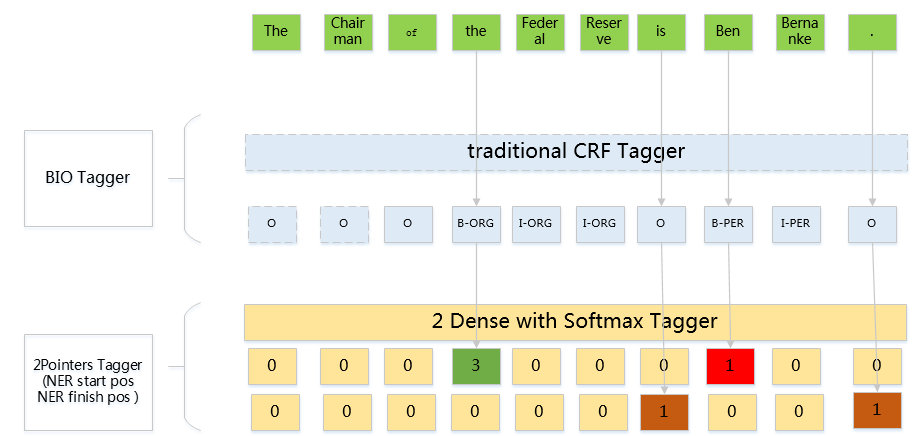
In this paper，我们继续发挥预训练模型和Attention的强项，基于Multilingual Bert和完全的Attention来优化Binlingual NER挖掘。本文的贡献在于：

1. 从两个单独的NER模型变成一个统一的双语模型。因为Multilingual BERT, 支持140种语言，基于双语语料生成一个统一模型，一方面是简化模型，防止过拟合，另一方面，也是一种数据增强方式。
2. Self Attention for tagging, No CRF. 因为M-BERT本身具备token embedding, position embedding and sentence embedding 能力，我们尝试用self attention 实现sequence tagging，而单语约束，可以利用sequence的softmax实现。
3. Cross Attention for Alignment。M-BERT 基于大量wiki和books语料，已经在多语言transfer体现效果。与SMT的Alignment是所有词的对齐不同，我们只考虑实体词对齐，将实体词对齐和NER同时训练，来满足bilingual constraints.

本文组织结构如下：第2部分是相关工作对比，重点对比NER的SOTA模型BiLSTM和META NET, 第3部分是META NET的理论分析，包括Embedding目标，Attention Tagger目标和Attention Align 目标， 第4部分是实验分析，包括Conll-2003英德、OntoNote 4英中的双语NER与SOTA的单语NER对比，还补充分析了双语NER一些性能不好的背后原因。第5部分是后续方向介绍。

2 Model Architecture

在对比介绍BiLSTM+CRF和META NET之前，我们先补充说明下从BIO标注，到2个Sequence Pointer的标注。

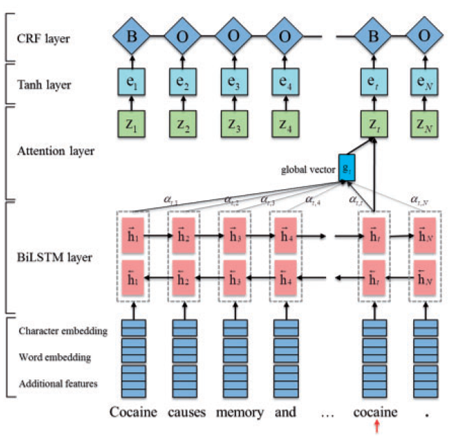


QA NET（）提出了2 Pointers Tag方法，相对于原来逐字标注某类型X的实体，B-X, I-X, O的标注模式，采用两组指针的标注模式更加简单清晰：

1. 头指针对于非实体开始(I-X, O)标注为0，对于实体开始，标注为1 （PER）, 2(LOC),3(ORG)，实际是一个多分类数组，如果不区分实体类别，头指针就是一个二分类数组。
2. 尾指针对于非实体结束标注为0，实体结束标注为1，其实就是一个二分类数组。
3. 实体的获取是从非0得Start Pointer找到最近的非0的End Pointer，即[Start pointer : End pointer], 然后根据Start Pointer值获取实体类型，即实现了多分类的实体标注。

NER 的SOTA模型是BiLSTM+CRF, 主要包括Embedding层、BiLSTM层，Attention层，Tanh和CRF层。其中:

1. Embedding层，主要是做字、词表示以及前后NGram的手工组合特征，目的是获得初级的语义描述。
2. BiLSTM层，则是通过双向LSTM做特征组合，可以学习长程的特征依赖。
3. Attention层，主要是SelfAttention，在长程特征依赖之上，能够通过注意力提升本句中字词的相关关系，更好学习字词的上下文特征。
4. Tanh和CRF层，通过非线性拟合和无向图模型来做Sequence的序列预测。如图2，总的来说，模型还是相对复杂，层数和不确定性也比较多。

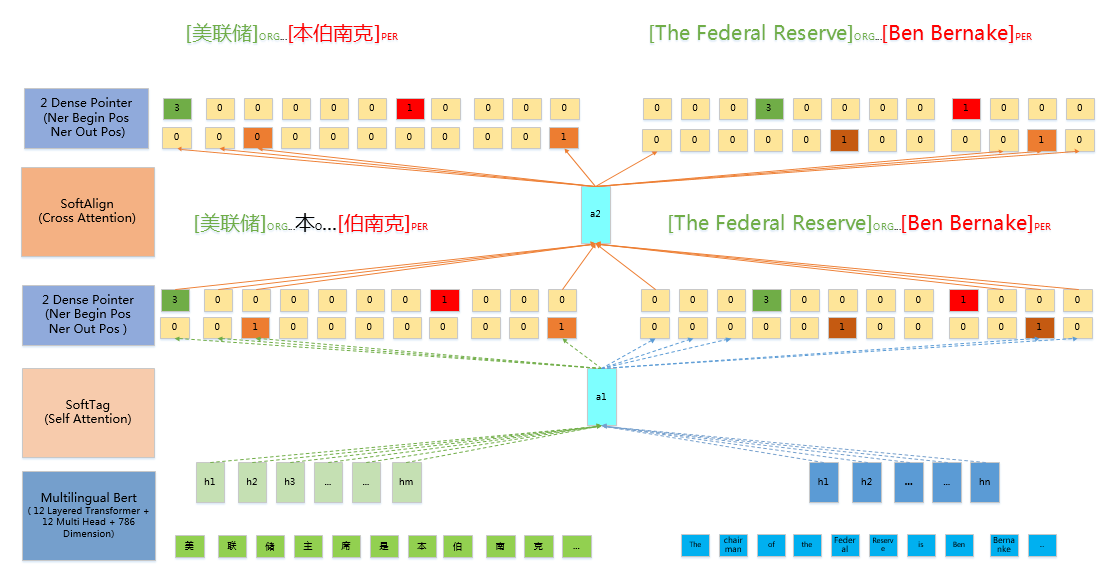


相对而言，META NET模型更加清晰和简单，如图3.

1. M-BERT替代多组手工特征。因为M-Bert本身具有token embedding, pos embedding和sentence embedding能力，就不需要单独做图2中表示层的三组手工特征。
2. Bilingual NER Tag粗调
   1. 从BiLSTM + Attention，简化为Attention。类似于（Attention is all you need）, 基于BERT的Attention能够很好的学习当前词的上下文特征，没有必要再做BiLSTM的时序特征组合。
   2. 从Tanh + CRF到2 Pointer Tagger。将传统的无向概率图模型CRF简化为两个多分类的数组，通过Start Pointer和End Pointer找到相应实体及实体类型。
   3. 对于双语NER，我们只需要一个模型，而不是两个模型。
3. Bilingual NER微调
   1. 增加SoftAlign, 通过Cross-Attention实现NER对齐，调整或补充NER，满足双语的NER约束。
   2. 增加2 Pointer Tagger, 类似于（3），在SoftAlign基础上，对NER进行微调，获取更好效果。

结合实际case，我们说明下具体过程：

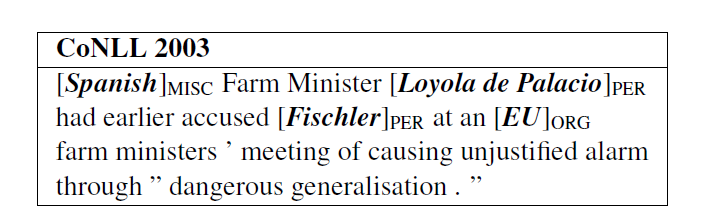
1. M-BERT层
   1. ” The chairman of the Federal Reserve is Ben Bernanke .” 被分词为 ['The', 'chairman', 'of', 'the', 'Federal', 'Reserve', 'is', 'Ben', 'Berna', '##nke', '.']
   2. “美联储主席是本伯南克。” 被分词为['美', '联', '储', '主', '席', '是', '本', '伯', '南', '克', '。']
2. SoftTag层
   1. 经过Self Attention 和 2 Pointer Tagger, 英文的实体标注为’[the Federal Reserve]ORG, [Ben Berna ###nke] PER’
   2. 而对应中文的实体标注‘[美 联 储]ORG, [本] O, [伯 南 克]PER’，
   3. 因为“本”不是中文的常见姓氏，所以对应的Start Tagger为0， 被识别为O
3. SoftAlign层
   1. 经过CrossAttention层，”the Federal Reserve”与“美 联 储”对齐，而“Ben Berna ###nke”与“本 伯 南 克”对齐
   2. 微调Ner Tagger标注， “本 伯 南 克”被标注为PER



3 Constraint and Algorithm

相当相关工作样，缺少跨语言的命名实体特征传递办法。

本文考虑：(1)将自有语料训练的词向量表示，改为基于跨语言预训练模型（Multilingual BERT），利用海量数据的Tokenizer将低频词分解为几个高频子词（SubWord）,降低UNK词比例。（2）基于SelfAttention + Dense的SoftTag，实现长距离的实体关注，利用长距离上下文提升命名实体的识别能力。（3）利用CrossAttention + Dense的SoftAlign,通过软实体对齐，将一种语言中的实体特征，传递到另外一种语言中，并且我们对实体窗口做适当的扩展，类似于一种fine-tuning效果，提升跨语言实体识别准确率。为了验证我们实际效果，我们在ConLL2003 增加WMT18的英德对齐语料，提升英语和德语的NER准确率，同时我们也尝试在OntoNotes 5.0增加WMT18的中英对齐语料，提升英语和中文的NER准确率，实验证明，3个单语的NER的准确率都有提升，均略有于单语的SOTA值，其中中文提升最为明显，提升超1.8%，在中文中，人名的提升也最为明显，提升超3.3%。



This is the second paragraph. It in formatted with the Text-indent style. Use of the Text-Indent style is required.

This Is a Subsubsection Heading. Use of This Style Sheet Is Required for all subsubsections.

This is example text. It is 10 point Times New Roman. This is example text. It is 10 point Times New Roman.

This is example indented text. It is 10 point Times New Roman. This is example indented text. It is 10 point Times New Roman. This is example indented text. It is 10 point Times New Roman. This is example indented text. It is 10 point Times New Roman.

This is an example of an extract or quotation. Note the indent on both sides. Quotation marks are not neces­sary if you offset the text in a block like this, and properly identify and cite the quotation in the text.

This Is an Example of a Figure Caption. Use of the figure caption style is required.

This Is a Section Heading   
This Style Sheet Must Be Applied   
to All Section Headings

Use of the section heading style is required.

Figures and Tables

This Is an Example of a Figure Caption.

You must have at least 6 points additional space between your figure or table and the figure or table caption. You must have at least 1 carriage return (12 points additional space) after the caption. You must also have at least 1 carriage return of space (12 points additional space) before a figure or table that does not begin at the top of a page. Figures and tables must not intrude into *any* margin or gutter (this includes boxes and rules and frames).

Margins

Papers **must** be printed in two-column format on 8-1⁄2 x 11 inch paper. The margins if you use 8-1/2 x 11 inch paper must be **exactly** as follows:

• Top margin: 3⁄4 of an inch

• Left margin: 3⁄4 inch

• Right margin: 3⁄4 inch

• Bottom margin: 1-1⁄4 inches

**Be certain that the margins apply to all pages in your document, not just the first section.** The actual image size for your paper (minus margins) is 7x 9 inches (17.78 x 22.86 cm). Papers that deviate from these measurements will not be published.

Style Sheet Use and Alterations

You are required to use the custom style sheets embedded in this document without alteration, including, but not limited to Paper Title, Author Name, Affiliation and Address, Abstract Head, Abstract Text, Section Heading, Text, Text-Indent, Subsection Heading, Subsubsection Heading, List Paragraph, Figure, and Figure Caption. Do not use the built-in Word styles such as Header, Normal, and Heading 1.

Fonts

All fonts must be embedded for your paper to be published. No exceptions.

Citation Style and Reference Style

All AAAI papers must use author year citations in the text (Engelmore and Morgan 1986). Numbered references may not be used. The references must be formatted using the exact references style shown the Author Formatting Instructions for Authors Using Word. Citations must be placed in alphabetical order. Lines must be justified. References may not be numbered. No indent is allowed. The style sheet will automatically add the proper amount of space between entries; no additional space should be added.

References

Clancey, W. J. 1984. Classification Problem Solving. In *Proceedings of the Fourth National Conference on Artifi­cial Intelligence,* 49-54. Menlo Park, Calif.: AAAI Press.

Engelmore, R., and Morgan, A. eds. 1986. *Blackboard Sys­tems.* Reading, Mass.: Addison-Wesley.

Hasling, D. W.; Clancey, W. J.; and Rennels, G. R. 1983. Strategic Explanations in Consultation. *The International Journal of Man-Machine Studies* 20(1): 3–19.

Rice, J. 1986. Poligon: A System for Parallel Problem Solving, Technical Report, KSL-86-19, Department of Computer Science, Stanford University, Stanford, CA.

Robinson, A. L. 1980a. New Ways to Make Microcircuits Smaller. *Science* 208:1019-1026.

1. Copyright © 2020, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved. [↑](#footnote-ref-1)