NER in Brazilian Portuguese

training a NER tagger for use in DOs

joão carabetta, lucas carneiro, bruno cuconato

EMAp

Named Entity Recognition for Brazilian Portuguese

overview

- we trained a neural network using Theano

Named Entity Recognition for Brazilian Portuguese

overview

- we trained a neural network using Theano
- we used as training and test sets the CoNLL-U files released by floresta sintá(c)tica, a Portuguese research project [1]

[1] http://www.linguateca.pt/floresta/corpus.html

Named Entity Recognition for Brazilian Portuguese

overview

- we trained a neural network using Theano
- we used as training and test sets the CoNLL-U files released by floresta sintá(c)tica, a Portuguese research project [1]
- the input corpus has some 10K sentences from Folha de S. Paulo and the Portuguese newspaper Público, fully revised by linguists

[1] http://www.linguateca.pt/floresta/corpus.html

motivation

the intent was to be able to recognize entities in Prefeitura de Mesquita's public register (Diário Oficial, D.O.)[2]

[2] http://transparencia.mesquita.rj.gov.br/

crawling PDF DOs

- crawled some 200 webpages for all the DOs published this year

DOs PDF

DIÁRIO OFICIAL
PREFEITURA MUNICIPAL DE MESQUITA
Mesquita, 12 de Abril de 2017 | Nº 00253.

Poder Executivo

JORGE MIRANDA
Prefeito
WALTINHO PAIXÃO
Vice-Prefeito

SUMÁRIO

DECRETO N.º 2059 DE 11 DE ABRIL DE 2017.

O PREFEITO MUNICIPAL DE MESQUITA, no uso de suas atribuições legais que lhe confere a Lei Nº. 994/16, de 04/11/16, publicada em 08/11/16, **DECRETA:**

Art. 1º - Abre Crédito Adicional Suplementar no

Art. 2º - Os recursos para atender a presente suplementação são oriundos da anulação total ou parcial das dotações abaixo relacionadas, conforme o exposto no inciso III, do artigo 43, da Lei nº. 4.320, de 17/03/64.

SECRETARIA MUNICIPAL DE GOVERNO, ADMINISTRAÇÃO E PLANEJAMENTO

PROGRAMA DE TRABALHO:

20.04.04.122.0143.2.153 – Despesas com Pessoal Decorrente da Contratação de Terceiros

ELEMENTO DE DESPESA:

Natureza da Despesa		Despesa	Fonte	Valor
3.3.90.34.00	Outros Pessoal Decor da	19	0	280.000,00

Figura 1: example Mesquita DO

extracting text from PDFs

- several tools available, none up to the task

extracting text from PDFs

- several tools available, none up to the task
- `pdftotext`: GNU/Linux shell utility

extracting text from PDFs

- several tools available, none up to the task
- `pdftotext`: GNU/Linux shell utility
  - fast

extracting text from PDFs

- several tools available, none up to the task
- `pdftotext`: GNU/Linux shell utility
  - fast
  - not very intelligent

Mesquita, 12 de Abril de 2017 | Nº 00253.

Poder Executivo
JORGE MIRANDA
Art. 2º - Os recursos para atender a presente suplementação são oriundos da anulação total ou parcial das dotações abaixo relacionadas, conforme o exposto no inciso III, do artigo 43, da Lei nº. 4.320, de 17/03/64.
Prefeito WALTINHO PAIXÃO Vice-Prefeito
SUMÁRIO SECRETARIA MUNICIPAL DE GOVERNO, ADMINISTRAÇÃO E PLANEJAMENTO

PROGRAMA DE TRABALHO:

extracting text from PDFs

- `PDFminer`: Python library

extracting text from PDFs

- `PDFminer`: Python library
  - more intelligent, but very slow

extracting text from PDFs

- `PDFminer`: Python library
  - more intelligent, but very slow
  - still not intelligent enough

preparing train/test set

the CoNLL-U format

CoNLL-U[3] files are plain text files (UTF-8, using only the LF character as line break) with three types of lines:

1. Word lines containing the annotation of a word/token in 10 fields separated by single tab characters; see below.

[3] http://universaldependencies.org/format.html

preparing train/test set

the CoNLL-U format

CoNLL-U[3] files are plain text files (UTF-8, using only the LF character as line break) with three types of lines:

1. Word lines containing the annotation of a word/token in 10 fields separated by single tab characters; see below.
2. Blank lines marking sentence boundaries.

[3] http://universaldependencies.org/format.html

preparing train/test set

the CoNLL-U format

CoNLL-U[3] files are plain text files (UTF-8, using only the LF character as line break) with three types of lines:

1. Word lines containing the annotation of a word/token in 10 fields separated by single tab characters; see below.
2. Blank lines marking sentence boundaries.
3. Comment lines starting with hash (#).

[3] http://universaldependencies.org/format.html

the CoNLL-U format

Sentences consist of one or more word lines, and word lines contain the following fields:

1. ID: Word index, integer starting at 1 for each new sentence; may be a range for multiword tokens; may be a decimal number for empty nodes.

the CoNLL-U format

Sentences consist of one or more word lines, and word lines contain the following fields:

1. ID: Word index, integer starting at 1 for each new sentence; may be a range for multiword tokens; may be a decimal number for empty nodes.
2. FORM: Word form or punctuation symbol.

the CoNLL-U format

Sentences consist of one or more word lines, and word lines contain the following fields:

1. ID: Word index, integer starting at 1 for each new sentence; may be a range for multiword tokens; may be a decimal number for empty nodes.
2. FORM: Word form or punctuation symbol.
3. LEMMA: Lemma or stem of word form.

the CoNLL-U format

Sentences consist of one or more word lines, and word lines contain the following fields:

1. ID: Word index, integer starting at 1 for each new sentence; may be a range for multiword tokens; may be a decimal number for empty nodes.
2. FORM: Word form or punctuation symbol.
3. LEMMA: Lemma or stem of word form.
4. UPOSTAG: Universal part-of-speech tag.

the CoNLL-U format

Sentences consist of one or more word lines, and word lines contain the following fields:

1. ID: Word index, integer starting at 1 for each new sentence; may be a range for multiword tokens; may be a decimal number for empty nodes.
2. FORM: Word form or punctuation symbol.
3. LEMMA: Lemma or stem of word form.
4. UPOSTAG: Universal part-of-speech tag.
5. XPOSTAG: Language-specific part-of-speech tag; underscore if not available.

the CoNLL-U format

6. FEATS: List of morphological features from the universal feature inventory or from a defined language-specific extension; underscore if not available.

the CoNLL-U format

6. FEATS: List of morphological features from the universal feature inventory or from a defined language-specific extension; underscore if not available.
7. HEAD: Head of the current word, which is either a value of ID or zero (0).

the CoNLL-U format

6. FEATS: List of morphological features from the universal feature inventory or from a defined language-specific extension; underscore if not available.
7. HEAD: Head of the current word, which is either a value of ID or zero (0).
8. DEPREL: Universal dependency relation to the HEAD (root iff HEAD $= 0$) or a defined language-specific subtype of one.

the CoNLL-U format

6. FEATS: List of morphological features from the universal feature inventory or from a defined language-specific extension; underscore if not available.
7. HEAD: Head of the current word, which is either a value of ID or zero (0).
8. DEPREL: Universal dependency relation to the HEAD (root iff HEAD $= 0$) or a defined language-specific subtype of one.
9. DEPS: Enhanced dependency graph in the form of a list of head-deprel pairs.

the CoNLL-U format

6. FEATS: List of morphological features from the universal feature inventory or from a defined language-specific extension; underscore if not available.
7. HEAD: Head of the current word, which is either a value of ID or zero (0).
8. DEPREL: Universal dependency relation to the HEAD (root iff HEAD $= 0$) or a defined language-specific subtype of one.
9. DEPS: Enhanced dependency graph in the form of a list of head-deprel pairs.
10. MISC: Any other annotation.

the CoNLL-U format

```
# text = PT no governo
# source = CETENFolha n=1 cad=Opinião sec=opi sem=94a
# sent_id = CF1-1
# id = 1
1   PT  PT  PROPN   PROP|M|S|@NPHR  Gender=Masc|Number=Sing
2-3 no  _   _   _   _   _   _   _
2   em  em  ADP <sam->|PRP|@N<  _   4   case    _   _
3   o   o   DET <-sam>|<artd>|ART|M|S|@>N   Definite=Def|Ge
4   governo governo NOUN    <np-def>|N|M|S|@P<  Gender=Masc
```

preparing train/test set

convert from CoNLL-U to two column format

```
«   O
Câmera  B-PROPN
Manchete    I-PROPN
»   O
é   O
o   O
nome    O
de  O
o   O
novo    O
programa    O
jornalístico    O
que O
estréia O
quarta-feira    O
.   O
```

preparing train/test set

problems

- POS tags did not discriminate the kind of proper noun found;

preparing train/test set

problems

- POS tags did not discriminate the kind of proper noun found;
- CoNLL-U is now (as of 2016) a well specified format, but in 2008 (when these files were made), there was still plenty of room for improvement

preparing train/test set

problems

- POS tags did not discriminate the kind of proper noun found;
- CoNLL-U is now (as of 2016) a well specified format, but in 2008 (when these files were made), there was still plenty of room for improvement
  - the input files had some inconsistencies that had to be corrected by hand

training a NN

- Name entity recognition with supervised learning

training a NN

- Name entity recognition with supervised learning
- It learns from a small corpora (as is common in linguistics)

training a NN

- Name entity recognition with supervised learning
- It learns from a small corpora (as is common in linguistics)
- Language independent

training a NN

- Name entity recognition with supervised learning
- It learns from a small corpora (as is common in linguistics)
- Language independent
- no resort to gazetteers

Name Evidence

- orthographic evidence (how does the name look like?)

Name Evidence

- orthographic evidence (how does the name look like?)
- distributional evidence (where does the word tend to occur?)

Model

- bidirectional LSTM with a sequential conditional random layer above it (LSTM-CRF)

Model

- bidirectional LSTM with a sequential conditional random layer above it (LSTM-CRF)
- Learns from recent inputs and preserve long-range dependencies CRF - Conditional Random Field

Model

- bidirectional LSTM with a sequential conditional random layer above it (LSTM-CRF)
- Learns from recent inputs and preserve long-range dependencies CRF - Conditional Random Field
- With the results of the LSTM, it tags the tokens according to its neighbourhood

Model

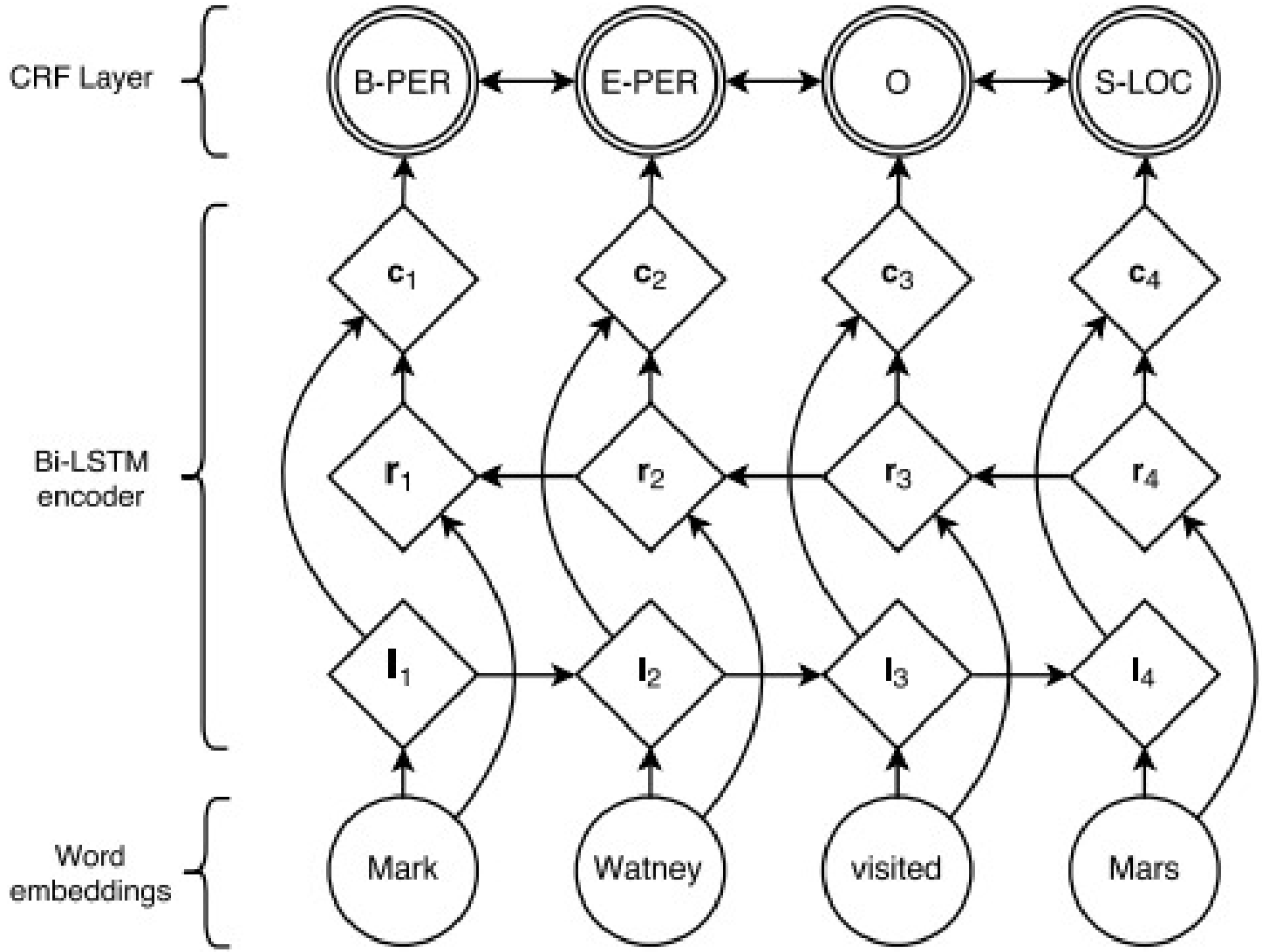

Figura 2: model schema

Model results

- accuracy: 99.06%

Model results

- accuracy: 99.06%
- precision: 94.23%

Model results

- accuracy: 99.06%
- precision: 94.23%
- recall: 92.49%

Model results

- accuracy: 99.06%
- precision: 94.23%
- recall: 92.49%
- F-score: 93.25%

testing on DO files

- because the extracted tests were not perfect, we manually built a test suite, with 35-odd *atos*

testing on DO files

- because the extracted tests were not perfect, we manually built a test suite, with 35-odd *atos*
- preprocessing

testing on DO files

- because the extracted tests were not perfect, we manually built a test suite, with 35-odd *atos*
- preprocessing
- we ran the NN on this test suite

testing on DO files

- raw input: several atos separated by ---

testing on DO files

- raw input: several atos separated by ---
- processed input is a sentence per line, one additional newline was added to separate one ato from another

testing on DO files

SECRETARIA MUNICIPAL DE EDUCAÇÃO PORTARIA Nº 394/2017 Dispõ
O SECRETÁRIO MUNICIPAL DE EDUCAÇÃO , no uso de suas atribu
RESOLVE : Art.1° Designar os servidores abaixo relacionado
· Isabelle da Cruz e Silva Guimarães - Mat.11/003.398-7 · A
Mat.11/008.117-5 Art.2° Esta portaria entra em vigor na dat
Thaís dos Santos Lima Secretário Municipal de Educação

RESOLUÇÃO CMDCA Nº 003 / 2017 .
Dispõe sobre a nova composição das Comissões de Trabalho do
O CONSELHO MUNICIPAL DOS DIREITOS DA CRIANÇA E DO ADOLESCEN
CMDCA , no uso das atribuições conferidas pela Lei Municipa
1º - Tornar público a nova composição das Comissões de Trab
Representantes da Sociedade Civil : Bruna Simãozinho Carval
Art .
2º - A Comissão de Políticas Públicas possui o acréscimo de
Representantes da Sociedade Civil : Bruna Simãozinho Carval
Comissão de Administração do Fundo Municipal Para a Infânci
COFIMIA : Representantes Governamentais : Fernandes de Mora

results

- output is `token__tag`

results

- output is `token__tag`
  - tag can be one of `B-PROPN`, `I-PROPN`, `O`

results

```
O__O PREFEITO__B-PROPN MUNICIPAL__I-PROPN DE__I-PROPN
MESQUITA__I-PROPN ,__O Estado__B-PROPN do__I-PROPN Rio__I-P
de__I-PROPN Janeiro__I-PROPN ,__O Sr.__O JORGE__B-PROPN
MIRANDA__I-PROPN ,__O no__O uso__O de__O suas__O atribuiçõe
lhe__O confere__O o__O art__O .__O 94__O ,__O IV__O ,__O da
Orgânica__B-PROPN Municipal__I-PROPN ,__O e__O nos__O termo
art__O .__O
```

results

- the NER tagger seems to rely too much on Capital Letters

results

- the NER tagger seems to rely too much on Capital Letters
- as it was trained on journalistic texts and not on DOs, it is not used to their extensive use of Capital Letters

results

```
CONVOCAÇÃO__B-PROPN PARA__I-PROPN REUNIÃO__I-PROPN ORDI
O__O CONSELHO__B-PROPN MUNICIPAL__I-PROPN DE__I-PROPN
ASSISTÊNCIA__I-PROPN SOCIAL__I-PROPN ,__O CONVOCA__B-PR
conselheiros__O titulares__O
```

next steps

- the model performs well even though the train set is very different from its use set

next steps

- the model performs well even though the train set is very different from its use set
- so the natural next step is to manually annotate DOs, so that the model can learn its idiosyncrasies

next steps

- the model performs well even though the train set is very different from its use set
- so the natural next step is to manually annotate DOs, so that the model can learn its idiosyncrasies
- as it was shown, the train set does not have to be very big, so it might be feasible

repository

Github