Labeling Multi Word Expressions with CRFs

Matthieu Constant (1), Isabelle Tellier (2)
Bachir Mounir Benmesbah (2)

(1) Université Paris-Est Marne-la-Vallée, LIGM
(2) université Sorbonne Nouvelle-Paris 3, Lattice

AIM-West 2015

Introduction

Motivations

- Use of CRFs (Conditional Random Fields)
- Combined with external resources
- For two tasks:
  - POS labeling joined with MWE identification for French
  - discontinuous (enventually embedded) MWE for English
- both treated as labeling tasks with the BIO convention

Introduction

Exemple for the First Task

“Max mange une pomme de terre.” (Max eats a potatoe)
becomes:
“Max/NPP mange/V une/DET pomme_de_terre/NC ./PONCT”
coded as:
“Max/NPP-B mange/V-B une/DET-B pomme/NC-B de/NC-I terre/NC-I ./PONCT”

Exemple for the Second task

“he was willing to budge a little on the price...”
becomes:
“he was willing to budge(1) a(2) little(2) on(1) the price...”
coded as:
“he was willing to budge/B a/b little/i on/I the price...”

Introduction

State of the Art

- Conditional Random Fields (CRFs) are discriminative probabilistic models generalising Maximum Entropy (MaxEnt)
- Combination with external ressources (Denis and Sagot, 2009, Constant and Tellier 2012)

Two Reference Corpora

- The leaves of the French Treebank, where MWE are identified
- An English corpus of User Generated Content with annotated discontinuous MWE (Schneider and alii 2014)

The task for both corpora

For the French task

- the tokens are the smallest possible ones
- POS labels will be associated with B (for Begin) or I (for In)
- MWE of various categories: verbal expressions (*fait l'objet, fait face*), P (*par rapport à*), "mots composés" NC (*eau de vie*), political functions (*chancelier de l'échiquier*), named entities (*New York*, *Banque de Chine*)... but no date!

For the English task

- the segmentation and the POS labels are provided
- possible discontinuity: B...O...O...I...I...
- two possible levels for MWE: B... (b...i...i)...I...I...

Example (reference corpus)

```
Quant_à P
la DET
technique NC
, PONCT
son DET
verdict NC
est V
implacable ADJ
. PONCT
```

Exemple (transformed into)

```
Quant P_B
à P_I
la DET_B
technique NC_B
, PONCT_B
son DET_B
verdict NC_B
est V_B
implacable ADJ_B
. PONCT_B
```

The French TreeBank (FTB) (Abeillé et al. 2003)

- Treebank transformed into POS labeled sentences
- version (Anne Abeillé 2005): 623 582 tokens, 569 080 “units” (single words, MWE, punctuations, numbers)
- simplified tagset: 29 POS labels (Crabbé et al, 2008)
- about 6% of “units” are MWE, 14% of the tokens belong to a MWE

Lexical Resources

Free Dictionarie

- DELA: general language (Courtois 1990, Courtois et al. 1997)
- Lefff: general language (Sagot 2010), extracted from MElt (Denis et Sagot 2009)
- Prolex: toponyms (Piton et al. 1999)
- Others: organisations (Martineau et al. 2009), first names (Unitex)

Local Grammars (Gross 1997)

- Name Entities: organisations, person names, etc. (Martineau et al. 2009)
- Dates (Blanc et al. 2007)
- Locative prepositions,
- Numbers as Déterminers (Constant 2003)

First Counts

- counts on 10% of the corpus
- Coverage: 1,5% of the units are unknown and out of the lexical resources (5% unknown units)
- Incomplete resources: 75,5% of MWE in the DEV are present in the lexical resources (87,5% including the training data)
- Incomplete Corpus: 30% of correct MWE from the resources found in the DEV are not annotated

Statistical Labeling

Principle

- Let x be a sequence of units $x = x_1 x_2 \ldots x_n$
- Let y be a sequence of labels $y = y_1 y_2 \ldots y_n$
- Purpose: find the best possible y^* knowing x

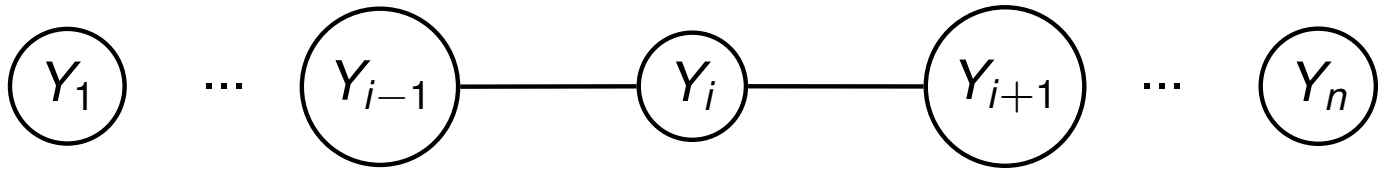
The Two Steps

- Train a model for $p(y|x)$ from examples (x, y), usually with maximum likelihood
- Annotate a new example x with this model:

$$y^* = argmax_y p(y|x)$$

by dynamic programming

Model Used for Linear CRFs



CRF Model [Lafferty et al. 01]

$$p(y|x) = \frac{1}{Z(x)} \prod_{c \in C} \exp\left(\sum_k \lambda_k f_k(y_c, x, c) \right)$$

- $Z(x)$: normalisation coefficient
- C: set of cliques of the dependency graph over Y
- y_c: value of the annotation on a clique $c \in C$
- f_k: feature fonctions (to be provided by the user)
- λ_k: weight of the f_k, parameters to be learned

Model Used for Linear CRFs

Training data :

x^1	x^2	x^3	y
Quant	1	0	P_B
à	0	0	P_I
la	0	0	DET_B
technique	0	0	NC_B
,	0	1	PONCT_B
...			

Definition of feature fonctions in CRF++/Wapiti

- A template = a ”card with holds” (or a colored shape)
- A feature function = a position of the template

Modèle des CRF linéaires

Training data :

x^1	x^2	x^3	y
Quant	1	0	P_B
à	0	0	P_I
la	0	0	DET_B
technique	0	0	NC_B
,	0	1	PONCT_B
...			

$f(x, y_c) = 1$ if (x^1=Quant) $\wedge$ (x^2=1) $\wedge$ (x^3=0) $\wedge$ (y=P_B)
$= 0$ elsewhere

Modèle des CRF linéaires

Training data :

x^1	x^2	x^3	y
Quant	1	0	P_B
à	0	0	P_I
la	0	0	DET_B
technique	0	0	NC_B
,	0	1	PONCT_B
...			

$f(x, y_c) = 1$ if $(x^1$=à$) \wedge (x^2$=0$) \wedge (x^3$=0$) \wedge (y$=P_I$)$

$= 0$ elsewhere

Model Used for Linear CRFs

Training data :

x^1	x^2	x^3	y
Quant	1	0	P_B
à	0	0	P_I
la	0	0	DET_B
technique	0	0	NC_B
,	0	1	PONCT_B
...			

$f(x, y_c) = 1$ if $(x^1$=la$) \wedge (x^2$=0$) \wedge (x^3$=0$) \wedge (y$=DET_B$)$
$= 0$ elsewhere

Model Used for Linear CRFs

Training data :

x^1	x^2	x^3	y
Quant	1	0	P_B
à	0	0	P_I
la	0	0	DET_B
technique	0	0	NC_B
,	0	1	PONCT_B
...			

$$f(x, y_c) = 1 \text{ if } (x^1\text{=Quant}) \wedge (x^2\text{=1}) \wedge (x^3\text{=0}) \wedge (y\text{=P\_B}) \wedge (x^1_{+1}\text{=à}) \wedge (x^2_{+1}\text{=0}) \wedge (x^3_{+1}\text{=0}) \wedge (y_{+1}\text{=P\_I})$$
$$= 0 \text{ elsewhere}$$

Model Used for Linear CRFs

Training data :

x^1	x^2	x^3	y
Quant	1	0	P_B
à	0	0	P_I
la	0	0	DET_B
technique	0	0	NC_B
,	0	1	PONCT_B
...			

$$f(x, y_c) = 1 \text{ if } (x^1\text{=à}) \wedge (x^2\text{=0}) \wedge (x^3\text{=0}) \wedge (y\text{=P\_I}) \wedge (x^1_{+1}\text{=la}) \wedge (x^2_{+1}\text{=0}) \wedge (x^3_{+1}\text{=0}) \wedge (y_{+1}\text{=DET\_B})$$
$$= 0 \text{ elsewhere}$$

Integrate knowledge into a CRF

Integrate inner properties (in columns $x^2, x^3 \ldots$)

- Starts by cap, contains a number, is a/contains a ponctuation, is the first word of a sentence, etc.
- Prefix or suffix of *n* characters

Integrate an external resource

- Filter *after the learning phase* with the same labels
  - only consider the best labeling compatible with the resource
  - two possible ways to operate: *a priori* or *a posteriori*
- Take into account *during the learning phase*
  - as new columns (still several ways to operate)
  - as new examples

Integrate an external resource

Examples of content of a resource

- quant à : *prep*
- la : *pronom*, *nom*, *det*
- technique : *nom*...

As a unique new column (labels can be either similar or different) :

			concatenated categories	
Quant	1	0	*prep_B*	P_B
à	0	0	*prep-prep_I*	P_I
la	0	0	*pronom-nom-det*	DET_B
technique	0	0	*nom*	NC
...				

Integrate an external resource

Examples of content of a resource

- quant à : *prep*
- la : *pronom*, *nom*, *det*
- technique : *nom*...

As new distinct columns (labels can be either similar or different) :

			distinct categories			
			prep	*pronom*	*nom*	*det*	
Quant	1	0	1	0	0	0	P_B
à	0	0	1	0	0	0	P_I
la	0	0	0	1	1	1	DET_B
technique	0	0	0	0	1	0	NC
...							

Integrate an external resource

Examples of content of a resource

- quant à : *prep*
- la : *pronom*, *nom*, *det*
- technique : *nom*...

As new examples (labels must be identical) :

quant	0	0	P_B
à	0	0	P_I
la	0	0	PRO
la	0	0	NC
la	0	0	DET

Labeling without/with MWE

- Filtering can bring a (little) improvement
- Taking into account an external resource *during the learning phase* significantly improves the results (97,33% versus 96,52% for POS only)
- As a single concatenated column (97,33% versus 96,6%)
- It is the same with MWE identification
- the segmentation (MWE identification) “costs” about 3% (best results: 94,42%)

Recognition of discontinuous MWE

	Schneider's results (perceptron)	our results (with CRFs)
without dictionnaries	55.57%	50.44%
with 2 dictionnaries	56.64%	55.23%
with 6 dictionnaries	61.95%	59.76%

Conclusion

Consolidation of known results

- CRFs still up to date..?
- easy to combine with external resource
- effective in condition to do it well
- also done for chunking...

MERCI
Questions?