ML challenges for MWE identification

Carlos Ramisch

PARSEME-FR meeting

January 15, 2018

Probabilistic NLP systems

Given an input X

- Enumerate all possible solutions

$$\mathcal{Y} = \{Y_1 \dots Y_n\}$$

- Weight all solutions according to their scores

$$p(Y_i)$$

$$p(Y_i) = ?$$

- Return the solution that maximises the score

$$\hat{Y} = \underset{Y_i \in \mathcal{Y}}{\operatorname{argmax}}\, p(Y_i)$$

Slide by Alexis Nasr

ML for probabilistic NLP

- Estimate the parameters of a score function $p(Y_i)$
  - Supervised learning: statistics over training data
  - Learn a function proportional to $p(X, Y_i)$ or $p(Y_i|X)$
- In NLP it is often impossible to observe/generate all possible solutions
  - All possible translations for a sentence
  - All possible POS-tag sequences
  - All possible syntax trees
  - All possible MWE identifications
  - ...
- To make problems treatable, the recipe is:
  - Decompose the problem into smaller pieces
  - Make independence assumptions
  - Use *clever* algorithms (dynamic programming, approximate search, etc.)

MWE Identification

- X is a sentence
  - *More often than not , however , it is not so straightforward to figure out how to make segmentation decisions , in order to split sentences into lexical units that make sense*
- $\hat{Y}$ is an annotation indicating where MWEs occur
  - **More often than not** , *however , it is not so straightforward to* **figure out** *how to* **make** *segmentation* **decisions** , **in order to** *split sentences into* **lexical units** *that* **make sense**
- What should $\hat{Y}$ look like?
  - span of tokens
  - set of pointers towards tokens
  - trees or graphs
  - pointers to lexicon entries
  - ...

ML and MWE identification

1. How to represent the MWE annotations on sentences?
2. How to decompose the problem into smaller pieces?
3. Which independence assumptions are reasonable?
4. What are the best algorithms to combine everything and solve the problem?

Things to take into account

Some characteristics of MWEs make them hard to identify

1. non-compositionality (idiosyncrasies)
2. discontinuities
3. ambiguity
4. nesting and overlap
5. variability
6. heterogeneity
7. rareness

Adapted from Constant et al. 2017

Challenge 1 : non-compositionality

MWEs are exceptions
The behavior of the whole is not predictable from the characteristics of the parts and of regular rules used to combine them

- Discovery techniques
- Word embeddings (Cordeiro's thesis)

Challenge 2 : discontinuities

- PARSEME shared task on verbal MWEs
- More frequent than one might initially think
- $\hat{Y}$ should be able to represent this (MWE = span vs. set of indices)

Challenge 3 : ambiguity

- Co-occurrence/structural ambiguity
  - *You promised to call me but you didn't,* **by the way***.*
  - *I recognize her* **by the way** *she walks*
  - *Je bois* **de la** *bière / Je parle* **de la** *bière*
- Semantic ambiguity
  - *The test was a* **piece of cake**
  - *I ate a* **piece of cake** *at the bakery*
- → Few expressions are highly ambiguous, most of them are not at all

Challenge 4 : nesting and overlap

- **Make plans** *and* **commitments**
- Quite rare, but it would be more elegant if we didn't ignore it completely

Challenge 5 : variability

- *He* **made a decision**
- *We are* **making** *a* **decision**
- *We* **make** *several hard and important* **decisions**
- *Important* **decisions** *should not be* **made** *hastily*
- *The* **decisions** *which we* **made** *yesterday*

Pasquer's thesis

Challenge 6 : heterogeneity

- *Multiword expressions* is a bad term
- "Distinct but related phenomena"
- Does it make sense to treat them uniformly?
- Learn several specialised models or one complex model?

Challenge 7 : rareness

- MWEs are frequent (the famous Jackendoff paper)
- Individual MWE categories are rare
- All depends on what you count as a MWE
  - Collocations?
  - Constructions?
  - Metaphors?
- Amount of training/test data to develop systems

Possible ML models

- Bag-of-words: classification-based approaches
- Sequences of words with Markov assumption: Sequence models
- Graphs: parsing-based methods

MWE identification and (deep) learning

- How to introduce more structure than sequence-based taggers?
  - To deal with overlap and nesting in a more principled way?
  - To deal with discontinuities in a more principled way?
- Since ambiguity is rare, should we use probabilistic models at all?
- What amount of training data allows us to make useful generalizations?
  - Deep learning requires large amounts of data
  - MWE annotation has made progress but are we there yet?
  - What is the impact of overfitting?
- Can word embeddings help predicting compositionality in context?
- Can we use character-based models to deal with variability?
- Can we perform discovery and identification at the same time to deal with rareness?