Outline

Multilingual NLP

1. Semantic Representations

2. Multilingual Resources

3. Projects
Semantic Representations

(Monolingual) Sentence Representations

- **Composition** of word embeddings using operations (+,×) on vectors and matrices

- Internal representations in **seq2seq** architectures or auto-encoders (NMT context vectors, skip-thought vectors...)

- **Latent paragraph vectors** in word2vec-like NNs
ML-NMT Context Vectors, why?

- Machine Translation is naturally a bilingual task
- Neural Machine Translation (NMT) encodes semantics in word vectors
- Straightforward extension of NMT to multilingual NMT (ML-NMT)
- ML word (or context) vectors lie in the same space
Semantic Representations in Neural Machine Translation Architectures

encoder

softmax

decoder

attention mechanism
Semantic Representations in Neural Machine Translation Architectures
The marshmallow has to be on top <eos>

Der Marshmallow muss
The marshmallow has to be on top

Der Marshmallow muss

Vocabulary
Marshmallow, muss, oben, drauf, sein...

Vocabulary
marshmallow, has, to, be, top...
Semantic Representations in Neural Machine Translation Architectures

Der Marshmallow muss oben drauf sein

< 2de > The marshmallow has to be on top
Semantic Representations in Neural Machine Translation Architectures

Der Marshmallow muss oben drauf sein

< 2de > De marshmallow moet er bovenop. < eos >
The marshmallow has to be on
The marshmallow has to be on

Vocabulary
Marshmallow, muss, oben, has, to, be, top...

Vocabulary
marshmallow, has, top, oben, bovenop, moet ...
ML-NMT: \( \{ \text{de}, \text{en}, \text{nl}, \text{it}, \text{ro} \} \rightarrow \{ \text{de}, \text{en}, \text{nl}, \text{it}, \text{ro} \} \)

2D t-SNT representation of context vectors
ML-NMT: \( \{ar, en, es\} \rightarrow \{ar, en, es\} \)

2D t-SNT representation of context vectors
## Semantic Textual Similarity Task (SemEval 2017)

<table>
<thead>
<tr>
<th>Model</th>
<th>track1 ar–ar</th>
<th>track2 ar–en</th>
<th>track3 es–es</th>
<th>track4a es–en</th>
<th>track5 en–en</th>
</tr>
</thead>
<tbody>
<tr>
<td>w2v 300-D</td>
<td>0.49</td>
<td>0.28</td>
<td>0.55</td>
<td>0.40</td>
<td>0.56</td>
</tr>
<tr>
<td>w2v 1024-D</td>
<td>0.51</td>
<td>0.33</td>
<td>0.59</td>
<td>0.45</td>
<td>0.60</td>
</tr>
<tr>
<td>ctx 1024-D</td>
<td><strong>0.59</strong></td>
<td><strong>0.44</strong></td>
<td><strong>0.78</strong></td>
<td><strong>0.49</strong></td>
<td><strong>0.76</strong></td>
</tr>
</tbody>
</table>

Pearson Correlation
The marshmallow has to be on top <eos>
Semantic Representations

in Neural Machine Translation Architectures II

$$\begin{align*}
\overrightarrow{h}_j &= \tanh\left( \overrightarrow{W} \|_{k=1}^{K} E_k x_{jk} + \overrightarrow{U} \overrightarrow{h}_{j-1} \right) \\
\overleftarrow{h}_j &= \tanh\left( \overleftarrow{W} \|_{k=1}^{K} E_k x_{jk} + \overleftarrow{U} \overleftarrow{h}_{j-1} \right)
\end{align*}$$
Semantic Representations

Multilinguality and Interlinguality through Factors

- Approximate phonetic encodings with **Metaphone 3**
- **Babel Synsets** for nouns (incl. named entities, foreign words and numerals), adjectives, adverbs and verbs. Negation particles are tagged with **NEG**
- **PoS, Lemma & Stem**
Semantic Representations

Multilinguality and Interlinguality through Factors

- Approximate phonetic encodings with Metaphone 3

- **Babel Synsets** for nouns (incl. named entities, foreign words and numerals), adjectives, adverbs and verbs. Negation particles are tagged with `NEG`

- **PoS, Lemma & Stem**

```
< 2de >|---| the|DET| the|0|-
marshmallow|NOUN| marshmallow|MRXML| bn:00053559n
has|VERB| has| HS | bn:00089240v to|PREP|to|T-
be|VERB| be| P  | bn:00083181v on|PREP|on|AN-
top|NOUN| top|TP | bn:00077607n
```
ML-NMT: \(\{de, en, nl, it, ro\} \rightarrow \{de, en, nl, it, ro\} \Leftrightarrow es, fr\)

2D t-SNT representation of context vectors
Semantic Representations

Improving NMT with Factors

Average for 20 language pairs
Semantic Representations

Beyond Zero-Shot NMT with Factors

en2fr

fr2en

en2es

es2en

BLEU_{system} - BLEU_{w}
Context vectors provide a better semantic representation at sentence level than the composition of word embeddings.

They are multilingual by construction.

Interlingual factors help to position in the same space new (related) languages.

... but you need time, GPUs and data.
Multilingual Resources

1. Semantic Representations

2. Multilingual Resources
   - BabelNet
   - Wikipedia

3. Projects
Data driven systems are (even) more prevalent with the boom of **deep learning** architectures.

Data **quality** is (even) more important for neural systems e.g. SMT vs. NMT.
Multilingual Resources

BabelNet

ウィキペディア
フリー百科事典
Multilingual Resources

BabelNet

- Multilingual encyclopedic dictionary
- Semantic network
- 271 languages
- 14 million entries
Multilingual Word Sense Disambiguation

de:  Es|- war|bn:00083181v ein|- riesiger|- Erfolg|bn:15350982n

en:  And|- it|- was|bn:00083181v a|- huge|bn:00098905a

success|bn:00075023n


nl:  En|- het|- was|bn:00083181v een|- groot|- succes|bn:06512571n

ro:  Ţi|bn:00012706n a|- fost|bn:00083181v un|- mare|bn:00098342a

succes|bn:00075024n
Multilingual Word Sense Disambiguation

de:  Es|- war|bn:00083181v ein|- riesiger|- Erfolg|bn:15350982n
en:  And|- it|- was|bn:00083181v a|- huge|bn:00098905a
success|bn:00075023n
nl:  En|- het|- was|bn:00083181v een|- groot|- succes|bn:06512571n
ro:  Şi|bn:00012706n a|- fost|bn:00083181v un|- mare|bn:00098342a
succes|bn:00075024n
Multilingual, web-based encyclopedia

- 299 languages
- 46 million articles

- *English*: 5,529,144 content articles;
  *Japanese*: 1,087,058 content articles
- Same article in multiple languages connected via interlanguage links

- Comparable corpora are easy to build
Same article in multiple languages connected via interlanguage links

Comparable corpora are easy to build

Categories allow to extract in-domain comparable corpora (not so easy!)

Parallel sentences can be extracted from comparable corpora (not so easy!)
In-domain Comparable Corpora

- Identify comparable articles

- Build a characteristic vocabulary for the domain of interest automatically from root articles

- Explore the Wikipedia categories’ graph to select the subset of categories in the domain
Graph Exploration

- Sport
  - Sports
  - Mountains
  - Mountains by country
  - Mountains of Andorra
    - Pyrenees
    - Mountains of the Pyrenees
  - Mountaineering
- Science
  - Scientific disciplines
  - Natural sciences
  - Earth sciences
  - Geology
  - Geology by country
  - Mountain ranges of Spain
  - Geology of Spain
Parallel Sentence Extraction

- Brute-force sentence-wise comparison for parallel pairs identification

- Affordable (sentences aligned at document level)

- Similarity measures to detect them
  - Cosine on character $n$-grams and pseudo-cognates, length factors
  - Cosine on context vectors
Projects

1. Semantic Representations

2. Multilingual Resources

3. Projects
   - CLuBS
   - QT21
   - WikiTailor
The CLuBS Project

CLuBS
Cross-Lingual Bibliographic Search

- **Initial Status**: Multilingual en/es/de/fr retrieval platform —PubPsych— online
- **Aim**: Make PubPsych cross-lingual and improve translation and retrieval with the latest advances
Projects

PubPsych – https://pubpsych.zpid.de/pubpsych/
Projects

The CLuBS Project

- German Project (Leibniz Gemeinschaft)
  - Multilingual NMT systems
  - In-domain parallel sentence extraction
  - Query translation?
  - Cross-lingual retrieval?
Projects

The QT21 Project

- **Aim**: Improve translation quality for all European languages

- **Initial Status**: 3 well covered languages, problems with morphologically complex and low-resourced languages
Projects

The QT21 Project

- European Project (Horizon 2020)
  - Multilingual NMT systems
  - Factored ML-NMT with interlingual information
Projects

WikiTailor

Your à-la-carte in-domain corpora extraction tool from Wikipedia

Joint work with Alberto Barrón-Cedeño
Projects

WikiTailor

Your à-la-carte in-domain corpora extraction tool from Wikipedia

Joint work with Alberto Barrón-Cedeño

- **Aim**: Extraction of parallel corpora in any domain and language from Wikipedia

- **Current Status**: Extraction of comparable corpora in any domain and language from Wikipedia
Thanks!