

System @IWLST2017

# Going Beyond Zero-Shot MT: Combining Phonological, Morphological and Semantic Factors

#### Cristina Espana-Bonet & Josef van Genabith

Our systems based on factored NMT are designed to fully exploit multilinguality by including factors that increase the number of common elements among languages such as phonetic coarse encodings and synsets, besides shallow part-of-speech tags, stems and lemmas. Document level information is also considered by including the topic of every document. This approach improves a baseline without any additional factor for all the language pairs and even allows beyond-zero-shot translation. That is, the translation from unseen languages is possible thanks to the common elements —especially synsets in our models— among languages.

### Factors

**Topic:** Biterm Topic Model learned for 20 topics, so that each topic is the main topic of at least 1% of the training documents

succes|NOUN|succes|SKSS|bn:06512571n|t2

M3 encoding: Approximate phonetic encoding with Metaphone 3

BN Synset: For nouns (incl. named entities, foreign words and numerals), adjectives, adverbs and verbs. Negation particles are tagged with NEG

**PoS:** {NOUN, VERB, PREPOSITION, PRONOUN, DETERMINER, ADVERB, ADJECTIVE, CONJUNCTION, ARTICLE, INTERJECTION } *Tool*: IXA pipeline & TreeTagger + Mappings

**Lemma:** *Tool*: IXA pipeline & TreeTagger

**Stem:** *Tool*: Snowball API

 $orient@@|VERB|orient|ARNTNK|bn:00091197v|t2 \quad ing|VERB|orient|ARNTNK|bn:00091197v|t2 \\$ 

(https://github.com/cristinae/BabelWE)

## Multilingual TED



< 2en > |-|-|-|+|t2 es|PRONOUN|es|AS|-|t2 war|VERB|war|AR|-|t2 ein|DETERMINER|ein|AN|-|t2riesiger|ADJECTIVE|riesig|RSJR|-|t2 Erfolg|NOUN|erfolg|ARFLK|bn:15350982n|t2 < 2de > |-|-|-|+|t2 and|CONJUNCTION|and|ANT|-|t2 it|-|it|AT|-|t2 was|VERB|was|AS|bn:00083181v|t2 a|DETERMINER|a|A|-|t2 huge|ADJECTIVE|huge|HJ|bn:00098905a|t2 success|NOUN|success|SKSS|bn:00075023n|t2 < 2en > |-|-|-|+|t2 ed|CONJUNCTION|ed|AT|-|t2 e|VERB|e|A|bn:00083181v|t2stato|VERB|stat|STT|bn:00083181v|t2 un|DETERMINER|un|AN|-|t2 enorme|ADJECTIVE|enorm|ANRM|bn:00102268a|t2 successo|NOUN|success|SKSS|bn:00078365n|t2 < 2en > |-|-|-|+|t2 en|CONJUNCTION|en|AN|-|t2 het|PRONOUN|het|HT|-|t2was|VERB|was|AS|bn:00083181v|t2 een|ARTICLE|een|AN|-|t2 groot|-|grot|KRT|-|t2

- ➡ High intersection of factors among all languages
- ◆ Tokens and lemmas are more similar in West Germanic languages, Babel synsets in Latin languages
- $\clubsuit$  Location in the ML-space is not improved by factors
- ◆ Further improvements: ML-WSD for synsets, PoS selection for M3 encodings, etc

# Architecture



 $<2en>|-|-|-|t2 \ si|NOUN|si|-|bn:00012706n|t2 \ a|VERB|a|A|-|t2 \ fost|VERB|fost|FST|bn:00083181v|t2 \ un|ARTICLE|un|AN|-|t2 \ mare|ADJECTIVE|mar|MR|bn:00098342a|t2 \ succes|NOUN|succes|SKSS|bn:00075024n|t2$ 

- S1: Several years ago here at TED, Peter Sk@@ill@@man introduced a design challenge called the marshmallow challenge. S1
- S2: And the idea's pretty simple: teams of four have to build the tallest free-@@standing structure out of 20 sticks
  - of spaghetti, one yard of tape, one yard of string and a marshmallow. S2
    - S3: The marshmallow has to be on top. S3
  - **S4:** And, though it seems really simple, it's actually pretty hard because it forces people to collaborate very quickly. **S4** 
    - **S5:** And so, I thought this was an interesting idea, and I incorporated it into a design workshop. **S5** 
      - S6: And it was a huge success. S6

S7: And since then, I 've conducted about 70 design workshops across the world with students and designers and architects, even the C@@T@@Os of the Fortune 50, and there's something about this exercise that reveals very deep lessons about the nature of collaboration, and I'd like to share some of them with you. S7

**S8:** So, normally, most people begin by orient@@ing themselves to the task. **S8** 





# **Zero-Shot Training Condition**



## **Small Training Condition**



This work was partially funded by the Leibniz Gemeinschaft (SAW-2016-ZPID-2), the BMBF through the project ALL SIDES (01IW14002) and the European Unions Horizon 2020 grant agreement No. 645452 (QT21).