Addressing Noise in Multidialectal Word Embeddings

Alexander Erdmann, Nasser Zalmout, and Nizar Habash
Computational Approaches to Modeling Language (CAMeL) Lab
New York University Abu Dhabi, Abu Dhabi, UAE
{ae1541, nasser.zalmout, nizar.habash}@nyu.edu

Introduction

- We compare several methods for learning Dialectal Arabic (DA) word embeddings via bidialectal dictionary induction in Maghrebi (Mag), Egyptian (Egy), Levantine (Lev), and Gulf (Glf)
- DA word embeddings are typically noisy due to:
  (a) Linguistic variation

Rabat	Cairo	Beirut	Doha	MSA	Gloss
مطيشة *mTyšħ*	قوطة *qwTħ*	بندورة *bndwrħ*	طماطم *TmATm*	طماطم *TmATm*	*tomato*
طبلة *Tblħ*	طربيزة *Trbyzħ*	طاولة *TAwlħ*	طاولة *TAwlħ*	مائدة *mAŷdħ*	*table*
لديد *ldyd*	حلو *Hlw*	طيب *Tyb*	لذيذ *lðyð*	لذيذ *lðyð*	*declicious*

  (b) Scarcity of corpora
  (c) Unstandardized orthography
  (d) Morphological complexity
  (a)–(d) reduce type frequencies causing data sparsity

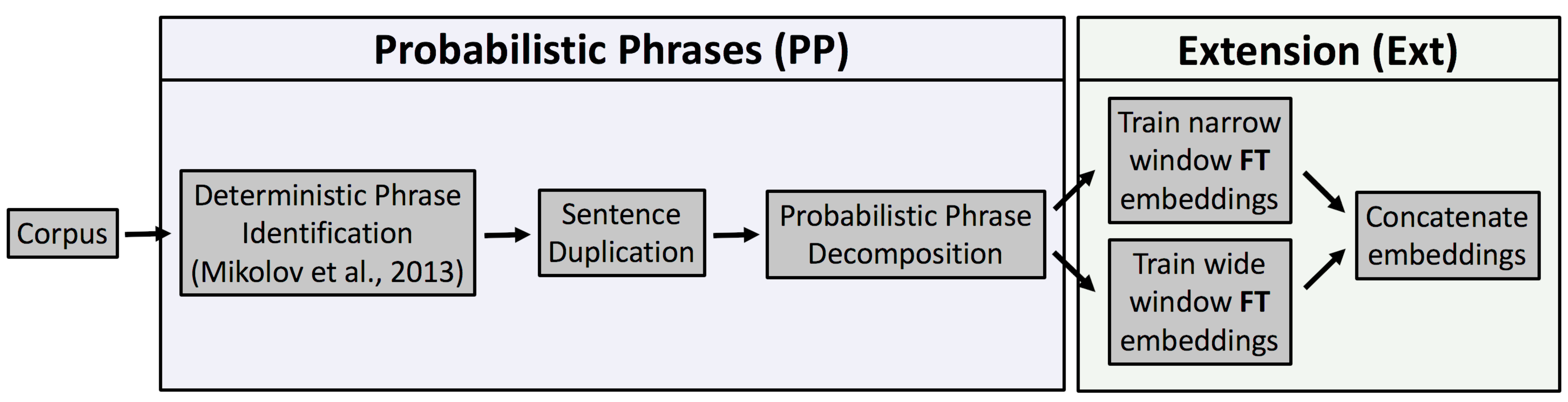
	DA-Egyptian	DA-Levantine	MSA	English
Tokens per type	20	19	68	128
Tokens with type frequency < 5	6%	6%	2%	1%

  (e) Orthographic ambiguity

- We target this noise with adaptations to the training pipeline that boost performance 2–53% in bi-dialectal dictionary mapping
- Most improvement is on low frequency forms, though high frequency forms improve slightly as well

Word Embedding Models

- Baseline **Fasttext** (FT) (Bojanowski et al., 2016)
  Incorporates subword information to model morphology
- **Extension** (Ext)
  FT vectors using narrow and wide context windows concatenated to model both syntactic and semantic similarities
- **Probabilistic Phrases with Extension** (PP+Ext)
  Ext vectors trained on multiple perspectives of each sentence generated by randomly joining/separating phrases



Systems for Representing Dialects in Common Space

Baseline representation
- **Identity** (ID)
  Maps every word to itself; metric of dialect similarity

Single embedding model trained on combined DA corpora
- **All Dialectal Arabic** (AllDA)
  One vector learned per type based on usage in all dialects

Dialect-specific models mapped into the same embedding space
- **Supervised Vecmap** (Svecmap) (Artetxe et al., 2016; 2017)
  Mapping leverages an iteratively augmented seed dictionary
- **Unsupervised MUSE** (Umuse) (Conneau et al., 2017)
  Mapping leverages adversarial training

Bidialectal Dictionary Induction Experiments

Metrics
- **Precision at *k* = 1** (P@1)
  Proportion of source words for which the nearest neighbor in the target dialect is a legitimate translation
- **Recall at *k* = 5** (R@5)
  Per-source-word average of the proportion of possible translations recalled in the nearest 5 target dialect neighbors
- **Weighted Recall at *k* = 5** (WR@5)
  R@5 weighted by source word type frequencies

Results

			SVECMAP			ALLDA	UMUSE
	Metric	**ID**	**FT**	**EXT**	**PP+EXT**	**PP+EXT**	**PP+EXT**
MAG	WR@5	28.9	35.3	42.2	**47.0**	32.6	26.8
⇓	R@5	24.9	36.2	40.4	**51.1**	26.2	14.9
LEV	P@1	33.6	35.3	39.7	**54.0**	33.7	12.2
MAG	WR@5	37.5	46.9	49.7	**50.8**	40.5	42.3
⇓	R@5	30.4	36.9	41.2	**45.2**	29.0	25.4
GLF	P@1	35.0	31.1	37.9	**40.0**	29.6	19.1
MAG	WR@5	42.4	48.2	**48.3**	47.9	45.8	43.1
⇓	R@5	30.7	34.5	39.4	**42.9**	34.0	25.5
EGY	P@1	36.0	29.4	**38.0**	36.6	36.3	20.9
EGY	WR@5	42.9	51.3	51.3	**52.8**	47.8	40.5
⇓	R@5	40.9	48.2	49.9	**52.8**	38.4	33.1
GLF	P@1	47.7	43.3	**48.5**	48.3	41.7	24.0
LEV	WR@5	43.2	50.6	50.4	**51.7**	48.5	40.9
⇓	R@5	33.6	37.8	38.9	**46.4**	31.8	24.7
GLF	P@1	39.0	34.1	37.5	**41.7**	33.1	20.0
LEV	WR@5	44.0	50.3	49.8	**52.4**	50.6	48.1
⇓	R@5	33.0	27.6	39.6	**42.3**	36.5	31.1
EGY	P@1	**39.6**	33.8	38.8	37.7	39.2	25.9

Findings

- PP+Ext generally outperforms other models according to all metrics, though improvement per R@5 is usually greater than per frequency weighted WR@5. This suggests **PP+Ext improves on infrequent words without compromising performance on frequent ones.**
- Noisy P@1 results suggest the **standard metric in the literature is not the most informative**. High polysemy caused by DA's orthographic ambiguity makes recall metrics more stable.
- Supervised mapping approaches outperform AllDA which outperforms unsupervised mapping. Yet, in less noisy environments, Artetxe et al. (2017) and Conneau et al. (2017) report the same unsupervised mapping approaches to rival the performance of the supervised approaches. Hence, **results achieved by imposing bilingual data scarcity constraints on non-noisy or monolingually rich environments may not generalize to truly low resourced, noisy, monolingually scarce environments such as DA**.