"Wilkommen, Salvete, Bienvenido" - No Problems... I know them all!

Srivatsan Srinivasan[1], Andrea Porelli[1], Ginevra Terenghi[2], Alessandro Bianchi[2]

1 - IACS, Harvard University, 2 - Politecnico Di Milano

Objectives

The project develops two models which transfer language understanding, for cross-lingual text classification - an integral component in measuring social media engagement of Tribe Dynamics' clients.

- Single classifier on a dynamically trained and aligned word embedding space.
- Single source Model Translation via learning a mixture of word translations.

Introduction

Tribe Dynamics provides in depth analysis of EMV (Earned Media Value) - dollar value estimate of social media marketing for cosmetic clients. To do so, they need a classifier model that suggests the relevance of a social media post to a brand. With its rapid expansion across continents and data labeling being done at a premium, Tribe Dynamics wishes to build cross-lingual classification models that are data-efficient and maintenance friendly.

Data and Current Models

- 106 cosmetic brands, $<$ 10000 data-points for each brand (labeled and unlabeled), 20+ languages (Italian for proof of concept)
- **Concerns:** Class imbalance, ground truth unreliability (10-15% mistakes in non-English)
- **Current Model:** Logistic Regression on n-grams joint vocabulary
- **Issues:** Scalability, sparsity, cost of data labeling, lack of transfer across languages, language differences (German, Asian, etc.)

Metrics	English	Others
AUC	0.97	0.95
PRC	0.92	0.71
F1	0.84	0.68

Table 1: Recreation of Tribe Dynamics' current baselines using Logistic Regression Classifier.

Classifier on Aligned Word Embeddings (CAWE)

We leverage pre-trained FastText MUSE embeddings(trained on Wikipedia) and dynamically train them further on our dataset using language model to capture the nuances of fashion social media vocabulary.

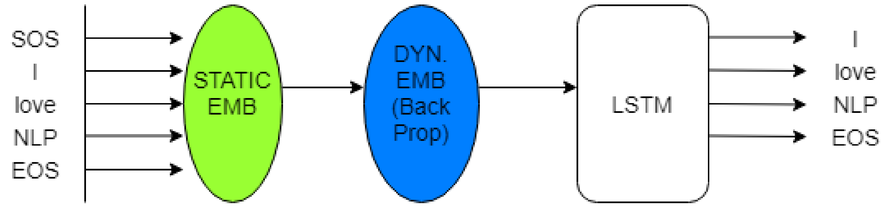

Figure 1: Language Model used to train Word Embeddings

- Embeddings map sparse binary vector (one-hot) to continuous D-dimensional space, i.e. $\mathbb{E} : [0,1]^{|V|} \rightarrow \mathbb{R}^D$.
- Language Model: Sentence $S \rightarrow w[1:T]$; Given $w[1:t]$ predict $w[t+1]$ as probability over vocabulary V.
- Encode similarity with proximity in embedded space since similar semantic words lead to similar successors.

Key Takeaways

CAWE

- Knowledge transfer through pre-training and alignment
- Similar neighbors, Classifier flexibility
- Single model, Compact representation.

MWT

- Knowledge transfer through word translation (bilingual lexicon) mixtures
- Interpretability, Classifier flexibility
- Single model, Saves data labeling drastically

Alignment of Embeddings

We align anchor words in both languages(s , t) using distance-preserving isometric transformations.

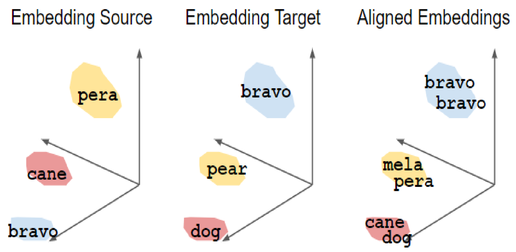

Figure 2: Similar semantics across languages helps alignment

$$\begin{aligned} &\text{Anchor Words - } \mathbb{A}, \quad \mathbb{X} = E_s(\mathbb{A}), \quad \mathbb{Y} = E_t(\mathbb{A}) \\ &W^* = \underset{W \in O_d(\mathbb{R})}{\arg\min} ||W\mathbb{X} - \mathbb{Y}||_F \quad \text{(Procrustes)} \\ &= XY^T = UV^T \quad \text{SVD} \end{aligned} \tag{1}$$

Results

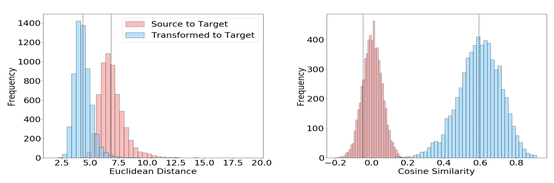

Figure 3: Embeddings before and after alignment. Transformed embeddings are closer to the target than the source.

Metrics	LR	CAWE
AUC	0.88	0.86
Avg. Prec.	0.95	0.95
F1	0.96	0.94

Table 2: CAWE Results(EN+IT). Comparable performance on all metrics with limited training on subset of data.

Mixture of Word Translations (MWT)

Why not have a single model in English and translate the rest? Sentence translation - costly, word translation (lexicon) - needs disambiguation.

Ex : "Amo **Dove**", "**Dove** sei"

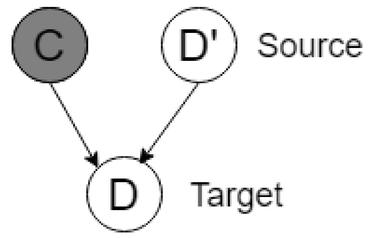

Figure 4: Graphical Model of generating a document

Generative model which treats a document as generating a bag of words. Needs only labeled data in source language (English) and unlabeled data in target (Italian). Labeled data in target needed for semi-supervised training.

$$\begin{aligned} P(d) &= \sum_c P(c) \sum_{d'} P(d|d',c) P(d'|c) \\ &= \sum_c P(c) \sum_{d'} \prod_{i=1}^{l} P(w_i|w_i',c) P(w_i',c) \end{aligned} \tag{2}$$

Expectation Maximization used to learn $P(w_i|w_i', c)$

$$\hat{C} = \arg\max_{c \in C} \prod_{i=1}^{V} \sum_{j=1}^{n_i} P(w_t^{ij}|w_s^i, c) s^{\lambda_{w_s^i} f_t(w_t^{ij}, c)} \tag{3}$$

- Interpretation: Model translation - through word translation - weighed by the probability of word translation.

Metrics	Baseline	MWT
AUC	0.83	0.86
Avg. Prec.	0.62	0.56
F1	0.80	0.74

Table 3: MWT Results(IT) on select brands(Dove, Kate, Vichy)