A High Coverage Method for Automatic **False Friends Detection** for Spanish and Portuguese

S. Castro, J. Bonanata y A. Rosá

Grupo de Procesamiento de Lenguaje Natural
Universidad de la República — Uruguay

VarDial 2018, COLING

Introduction

Objective: classify between false friends or cognates for Spanish-Portuguese

False friends: pair of words from different languages that are written or pronounced in a similar way, but have different meanings.

Example False Friends

obligado — obrigado

no — no

aceite — aceite

borracha — borracha

cadera — cadeira

desenvolver — desenvolver

propina — propina

Motivation

False friends make harder to **learn a language** or to **communicate**, especially when it's similar to the mother tongue.

- Between Spanish and Portuguese, the amount of cognates reaches the 85% of the total vocabulary (Ulsh, 1971).

Related Work

Frunza, 2006: supervised machine learning using **orthographic distances** as features to classify between **cognates**, **false friends** or **unrelated**.

Related Work

Mitkov et al., 2007: used a combination of distributional and taxonomy-based approaches. Worked with English-French, English-German and English-Spanish.

They use WordNet taxonomy similarities to classify, and if a word is missing they fall back to a distributional method.

Related Work

Mitkov et al., 2007

For the distributional method they build vectors based on word windows, computing the co-occurrence probability. Then, they compared the N closest words of each word in the pair, translate one of them and count occurrences in the other one. They defined a threshold based on Dice coefficient.

Related Work

Ljubešić et al., 2013: based on (Mitkov et al., 2007), experiment with several ways to build the vector space (e.g. tf-idf) and measure vector distances (e.g. cosine distance). They also proposed to use PMI.

They worked with closely related languages: Slovene and Croatian.

Related Work

Sepúlveda and Aluísio, 2011: false friends resolution for Spanish-Portuguese, highly based on (Frunza, 2006).

They added an experiment with a new feature whose value is the **likelihood of translation**, from a probabilistic dictionary (generated taking a large sentence-aligned bilingual corpus).

Word Vector Representations

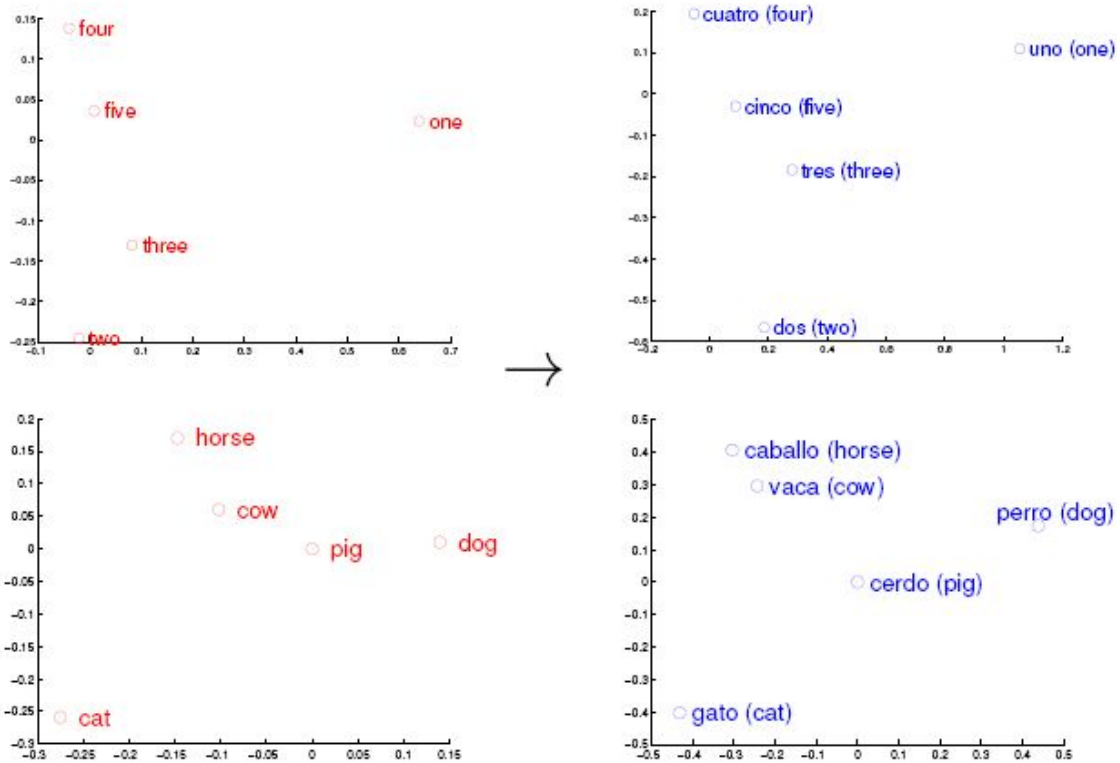
Related work crafted their own word vector representations. We propose to use the skip-gram-based **word2vec** model (Mikolov et al, 2013a).

Transform between Vector Spaces

Mikolov et al, 2013b: propose a method to correspond two word2vec vector spaces via a **linear transformation**.

Used to build dictionaries and phrase tables.

Transform between Vector Spaces

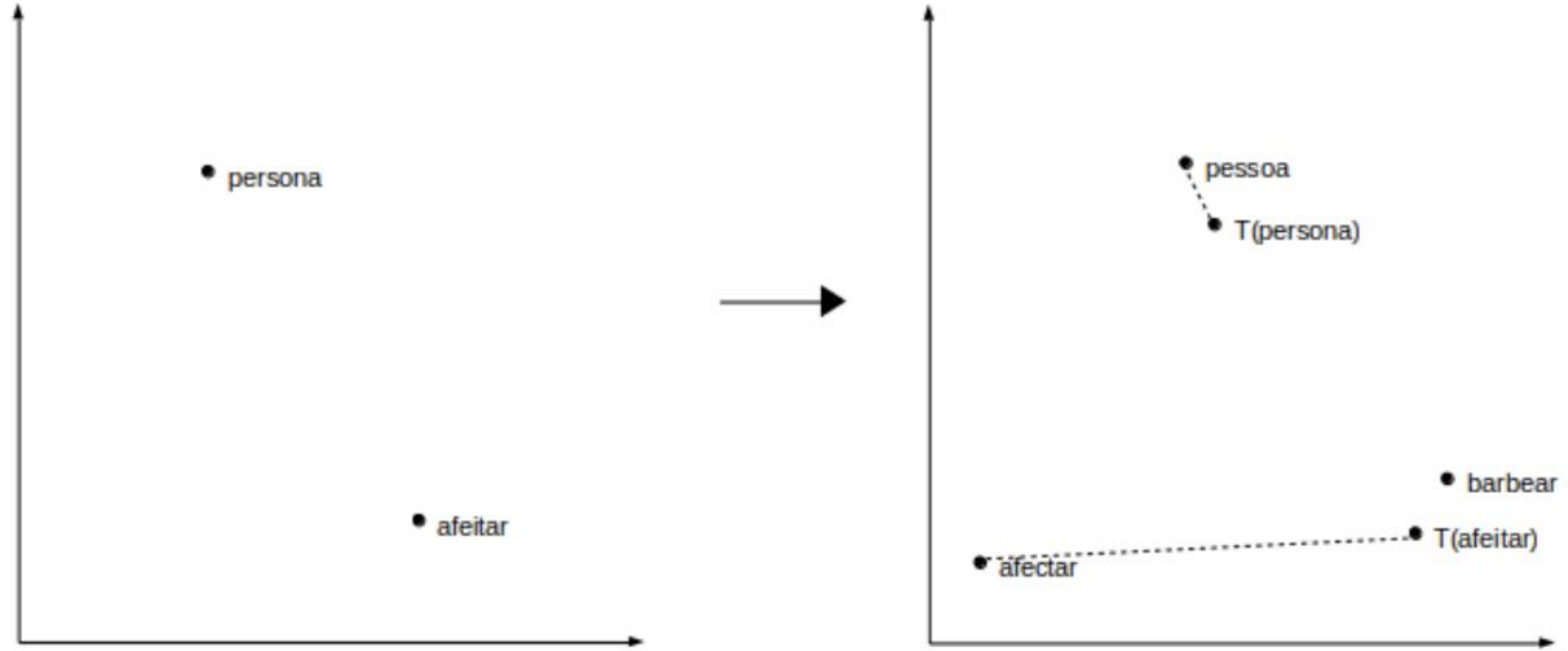


Our Method

Build word2vec vector spaces, find a linear transformation and measure vector distances.

Note that we don’t cope with related/unrelated, we just focus on cognate/false friends

Our Method


persona
afeitar
pessoa
T(persona)
barbear
T(afeitar)
afectar

Our Method

We used the **Wikipedia**’s for the vector spaces.

Open Multilingual WordNet (Bond and Paik, 2012) was used as a bilingual lexicon to fit the linear transformation: we iterate over synsets and took lexical units from each language. Then we employed Least Squares.

Our Method

We take one of the word vectors, transform it to the other space and compute:

1. The cosine **distance** between T(source_vector) and target_vector.
2. The **number of word vectors** in the target vector space **closer** to target_vector than T(source_vector).
3. The **sum of the distances** between target_vector and T(source_vector_i) for the **top 5** word vectors source_vector_i **nearest** to source_vector.

Experiments

We used (Sepúlveda and Aluísio, 2011) dataset, which is composed by 710 pairs (338 cognates and 372 false friends).

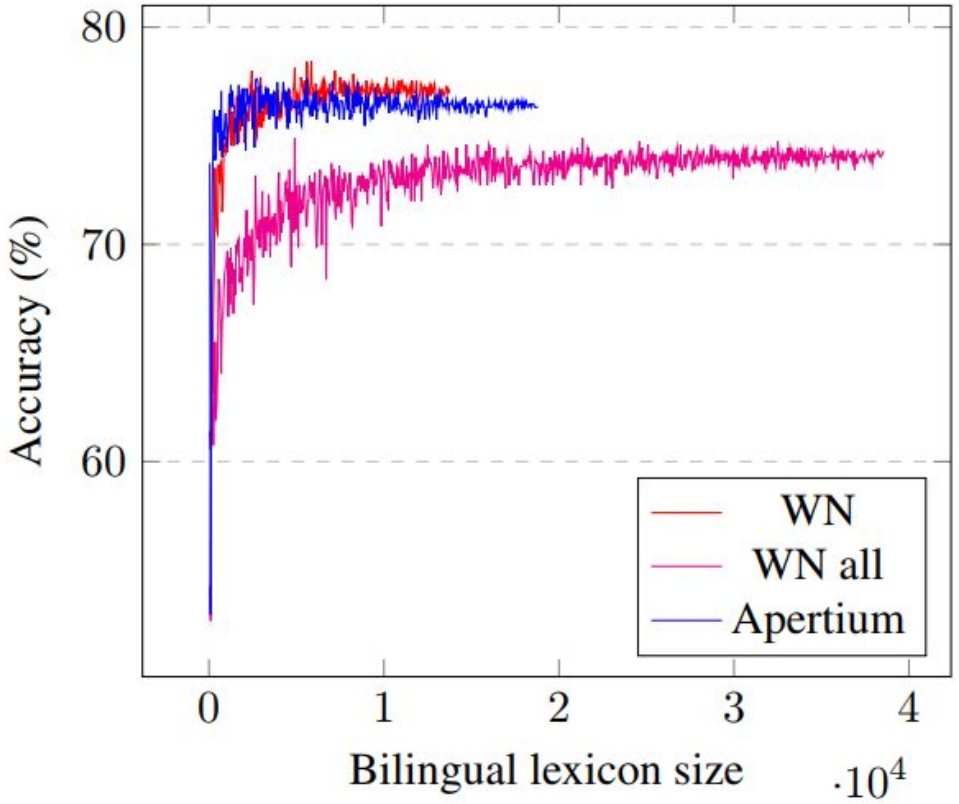
Experiments

Method	Accuracy	Coverage
WN Baseline	68.18	55.38
Sepúlveda 2	63.52	100.00
Sepúlveda 3.2	76.37	59.44
Apertium	77.75	66.01
Our method	77.28	97.91
+ frequencies	79.42	97.91

Experiments: different configurations

Method configuration	Accuracy
es-400-100	**77.28**
es-800-100	76.99
es-100-100	76.98
es-200-100	76.84
es-200-200	76.55
pt-200-200	76.13
es-200-800	75.99
pt-400-100	75.99
pt-100-100	75.84
es-100-200	75.83
es-100-100-2	74.98

Experiments: bilingual lexicon



Conclusions

- We have provided a new approach to classify false friends with high accuracy and coverage.
- We studied it for Spanish-Portuguese and provided state-of-the-art results for the pair.
- The method doesn’t require rich bilingual datasets.
  - It could be easily applied to other language pairs.

Future Work

- Experiment with other word vector representations and state-of-the-art vector space linear transformation.
- Work on fine-grained classifications.
  - E.g., partial false friends.

Thank you!

Questions?