Multimodal aspects of semantics

Experiment 1: Discovering the color of concrete objects

Universitat Pompeu Fabra
Barcelona

Oriol Jiménez, Roger González and Ana Caicoya

Assignment 2 - Computational Semantics

1. Introduction

Hypothesis

The semantic **relationship** between **nouns** (representing concrete entities) and their typical **colors** is more accurately captured by vector distances in **English** language models compared to **Spanish**.

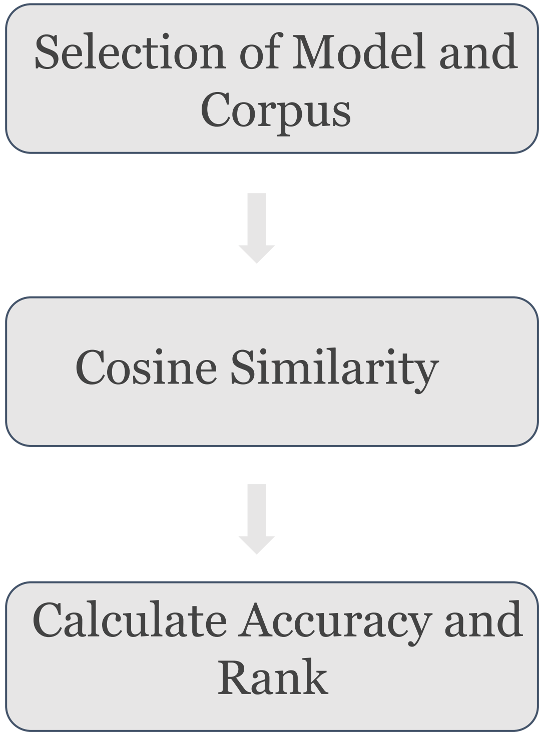
Dataset

52 Nouns - 11 Colors.

Examples

“crow–black”, “parsley–green”

Methodology



Selection of Model and Corpus

Cosine Similarity

Calculate Accuracy and Rank

! Ref: Bruni, E., G. Boleda, M. Baroni, N. K. Tran (2012), Distributional semantics in technicolor. Proceedings of ACL 2012, pp. 136-145, Jeju Island, Korea.

1. Data processing

English

- Compute cosine similarity with all colors
- Take the maximum
- Take the second maximum
- Calculate in which position is the correct colour

Spanish

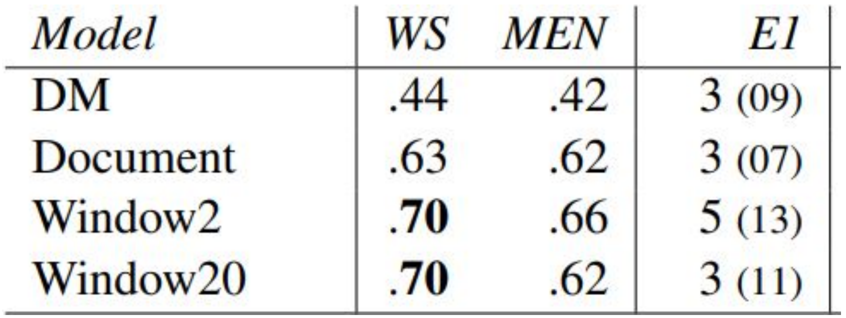
- Translate data.
- Follow same procedure as in English

Models

fastText word2vec
200000 words
English: cc.en.300.vec.gz
Spanish: cc.es.300.vec.gz

2. Comparing spanish and english results

Count Based Embeddings



Model	*WS*	*MEN*	*E1*
DM	.44	.42	3 (09)
Document	.63	.62	3 (07)
Window2	**.70**	.66	5 (13)
Window20	**.70**	.62	3 (11)

Table 1: Results from the Technicolor paper based on counts. The 'E1' column shows the mean rank, with the count of correct first positions noted in parentheses.

Modern embeddings

Model	Mean Rank (ENG)	Mean Rank (SPA)
fastText	1.5 (31)	2.0 (22)
USE	3.0 (18)	2.5 (19)
mBERT	5.0 (7)	5.0 (6)
roBERTa (base)	6.0 (2)	6.5 (9)
roBERTa (large)	4.0 (7)	5 (4)

Table 2: Comparison of results in Spanish and English. Results with some modern embeddings. Mean Ranks and count of right fist position noted in parentheses.

! Ref: Bruni, E., G. Boleda, M. Baroni, N. K. Tran (2012), Distributional semantics in technicolor. Proceedings of ACL 2012, pp. 136-145, Jeju Island, Korea.

! We have used xlm-roBERTa trained over 100 languages.

3. Comparing spanish and english results

1.
Overall accuracy

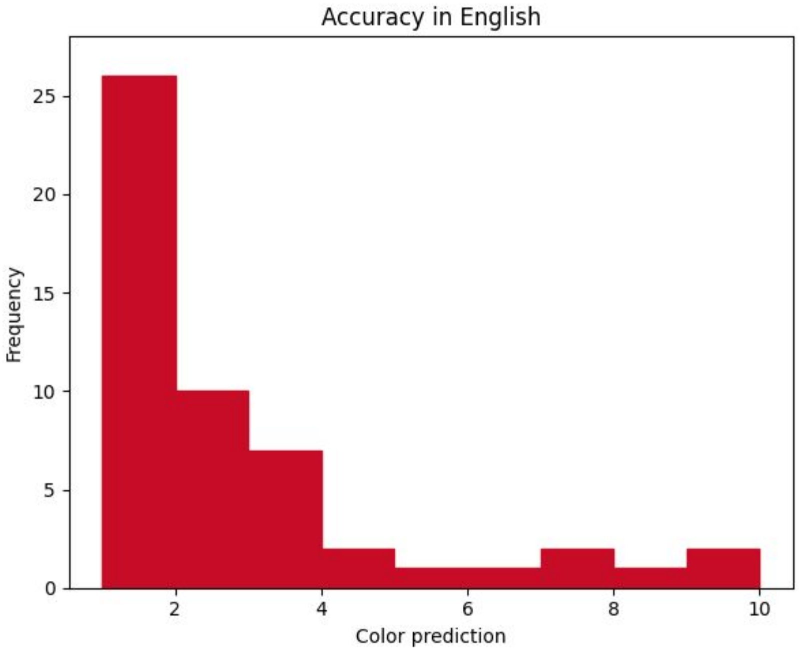
English: 50.00%
Spanish: 42.31%

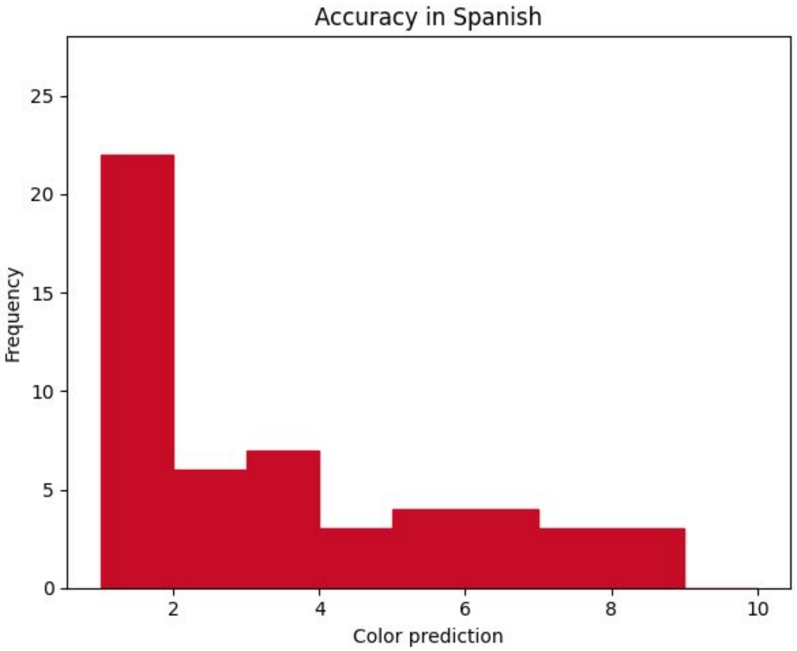
2.
Second accuracy

English: 69.23%
Spanish: 53.85%

3.
Cross-Linguistic Accuracy (fastText)

26.92%





4. Error analysis

Fruit association

Banana & cherry with color orange

High differences

“Pig” - “pink”
In Spanish it’s the 8th choice.

Gold Standard Ambiguity

Cloud with grey/azul (blue) or soil with brown/green

Agreement on common words

“blood”, “grass”

5.Conclusions

Cross-Linguistic Insights

English models may be benefiting from more **robust** linguistic **data** and **research**, potentially due to the language's global predominance.

Relationship Insights

Color **perceptions** embedded in English may lead to stronger noun-color associations due to **cultural** emphasis, can **affect** the **frequency** and context of colors.

Data and Model Influences

The corpus and **data quality** used for training language models are crucial. English language data might be more **extensive and varied**, leading to more nuanced color associations in models.