When the whole is greater than the sum of its parts: Multiword expressions and idiomaticity

Aline Villavicencio

University of Essex (UK)

Federal University of Rio Grande do Sul (Brazil)

Multiword Expressions

11 TV Shows That <u>Jumped The Shark</u>

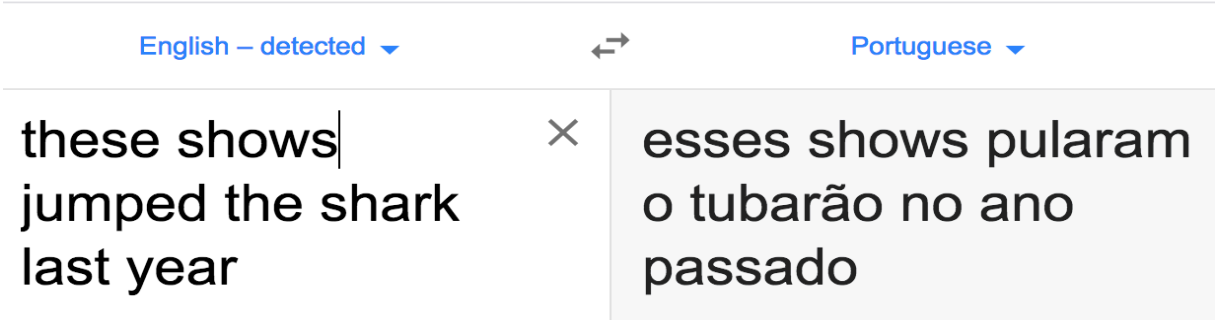
- Refers to the specific moment when a TV show goes downhill. Originally from *Happy Days*
- We may get lost in translation

Multiwords and NLP

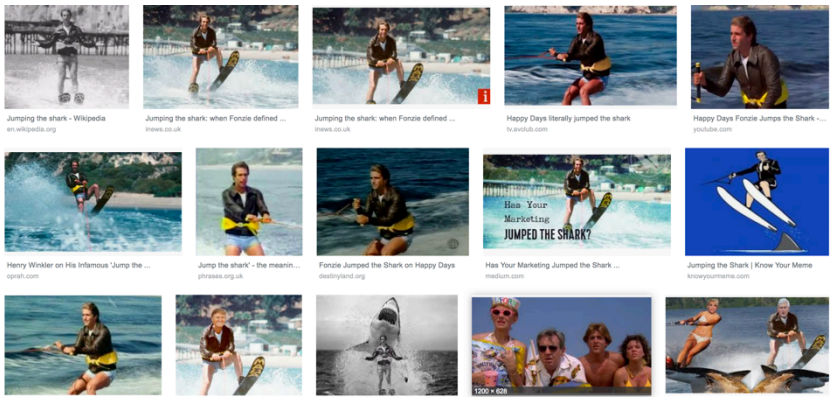
An open problem in NLP

(Schone and Jurafsky, 2001)

- Machine Translation



- Text Simplification
  - They moved over the fish
- Information Retrieval



Multiword Expressions (MWEs)

- **Recurrent** or typical combinations of words
  - That are **formulaic** (Wray 2002)
  - That need to be treated as a **unit** at some level of description (Calzolari et al. 2002)
  - Whose **interpretation crosses word boundaries** (Sag et al. 2002a)
- MWE Categories
  - **Verb-noun combinations**: *rock the boat, see stars*
  - **Verb-particle constructions**: *take off, clear up*
  - **Lexical bundles**: *I don't know whether*
  - **Compound Nouns**: *cheese knife, rocket science*

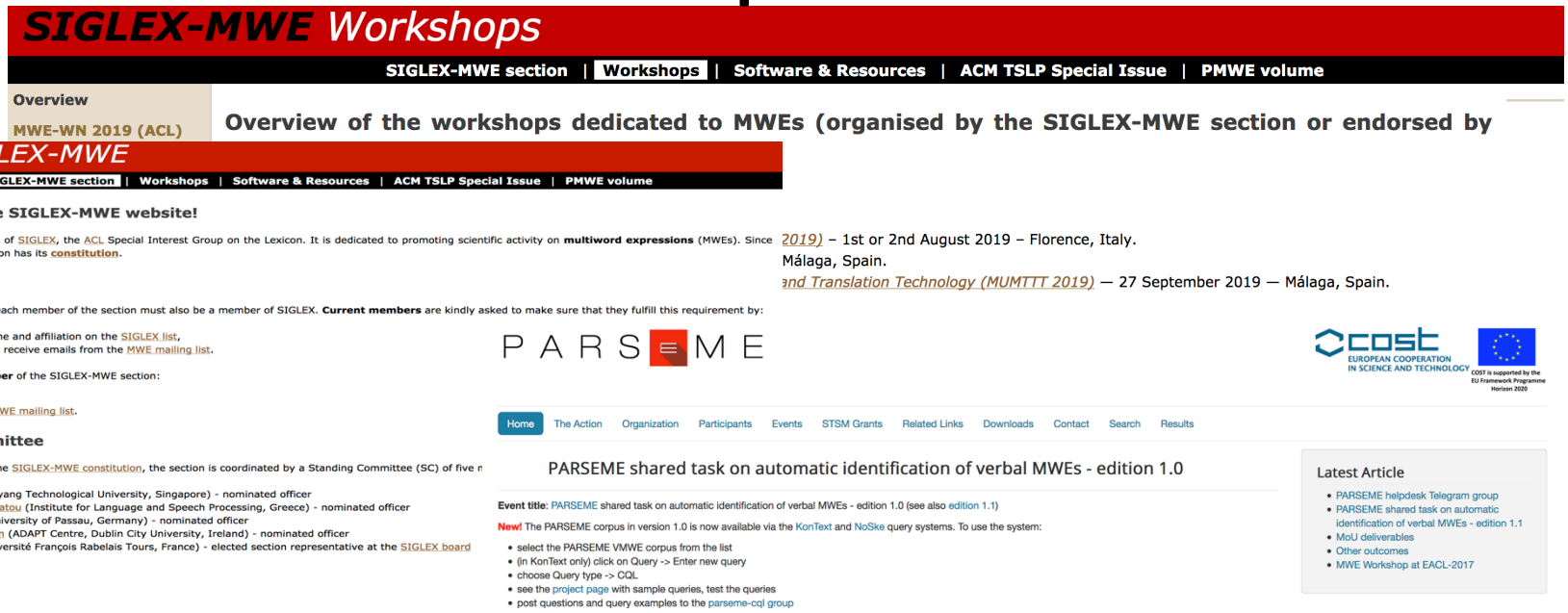
Multiword Expressions (MWEs)

- **High degree of lexicalisation**
  - *happy as a sandboy*
- **Breach of general syntactic rules/greater inflexibility**
  - *by and large/*short/*largest*
- **idiomaticity** or **reduced semantic compositionality**
  - *olive oil:* ***oil*** *made of* ***olive***
  - *trip the light fantastic: to dance*
- **high degree of conventionality and statistical markedness**
  - *Fish and chips, strong/?powerful tea*

MWEs are all around

- 4 MWEs produced per minute of discourse (Glucksberg 1989)
- Same order of magnitude in mental lexicon of native speakers (Jackendoff 1997)
- Large proportion of technical language (Biber et al. 1999)
- Faster processing times compared to non-MWEs (Cacciari and Tabossi 1988; Arnon and Snider 2010; Siyanova-Chanturia 2013)

Multiword Expressions



- 17 years and over 1000 citations after Sag et al. (2002) *Pain in the Neck* paper
- 16 years after the first MWE workshop and
- Many projects later

They are still an open problem

What's the big deal?

- MWEs come in all shapes, sizes and forms:
  - Idioms
    - *keep your breath to cool your porridge*
      - keep to your own affairs

  - Collocations
    - *fish and chips*

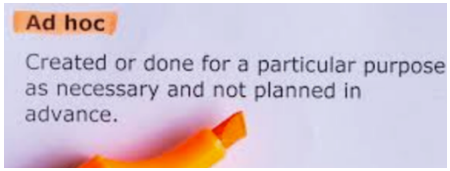
- Models designed for one MWE category may not be adequate for other categories

What's the big deal?

- MWEs may display various degrees of idiosyncrasy, including lexical, syntactic, semantic and statistical (Baldwin and Kim2010)
  - a *dark horse*

    - *colour of horse*
    - *an unknown candidate who unexpectedly succeeds*

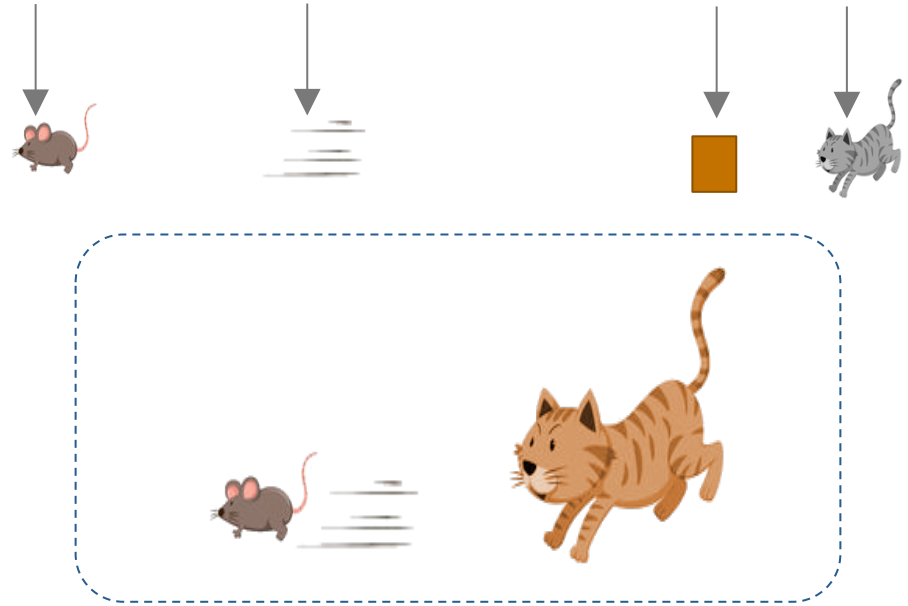
  - *ad hoc*
    - What is *hoc*?



  - To wine and dine
    - *wine* used as a verb

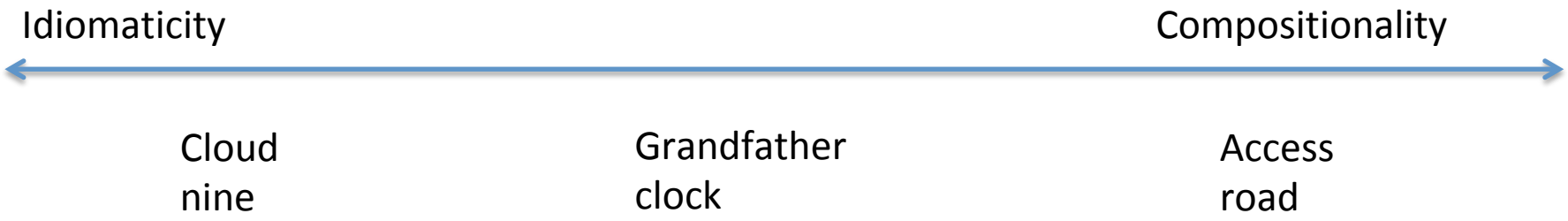
What's the big deal?

- NLP and Principle of Compositionality
  - The meaning of the **whole** comes from the meaning of the **parts**.
    - *"The mouse is running from the brown cat"*



What's the big deal?

- Meaning of MWE may not be understood from meaning of individual words
  - *brick wall* is a wall *made of* bricks,
  - *cheese knife* is not a *knife made of* cheese → *knife for cutting* cheese (Girju et al., 2005).
  - *Loan shark* is not a shark for loan but a person who offers **loans** at extremely high interest rates



In sum

- For NLP, given a combination of words determine if
  - It is a MWE
    - *Rocket science* vs. *small boy*
  - How syntactically flexible it is
    - *Kick the bucket, ?the bucket has been kicked*
  - If it is idiomatic
    - *Rocket science* vs. *olive oil*
- Decide if it can be processed accurately using Compositional Methods
    - *the meeting was cancelled as he kicked the bucket*
    - *a reunião foi cancelada quando ele chutou o balde*

In sum

- Clues from:
  - Collocational Properties
    - Recurrent word combinations
  - Contextual Preferences
    - (Dis)similarities between MWE and word part contexts
  - Canonical Form Preferences
    - Limited preference for expected variants
  - Multilingual Preferences
    - (A)symmetries for MWE in different languages

In this talk

- Collocational Properties
- Canonical Form Preferences
- Contextual Preferences
- Conclusions and Future Work

COLLOCATIONAL PREFERENCES

Collocational preferences

- Collocations of a word are statements of the habitual or customary places of that word (Firth 1957)
  - Statistical markedness detected by measures of association strength

Association Measures	
1. Pointwise Mutual Information (PMI)	$\log \frac{p(w_1 w_2)}{p(w_1 *)p(* w_2)} = \log \frac{f(w_1 w_2)}{f_\emptyset(w_1 w_2)}$
2. Specific Total Correlation (STC)	$\log \frac{p(w_1 w_2 w_3)}{p(w_1 **)p(* w_2 *)p(** w_3)} = \log \frac{f(w_1 w_2 w_3)}{f_\emptyset(w_1 w_2 w_3)}$
3. Specific Information Interaction (SII)	$\log \frac{p(w_1 w_2 *)p(* w_2 w_3)p(w_1 * w_3)}{p(w_1 **)p(* w_2 *)p(** w_3)p(w_1 w_2 w_3)}$
4. Students-t based association (t)	$\frac{f(w_1 \ldots w_n) - f_\emptyset(w_1 \ldots w_n)}{\sqrt{f(w_1 \ldots w_n)}}$
5. Dice	$\frac{n\ f(w_1 \ldots w_n)}{f(w_1) + \ldots + f(w_n)}$
6. χ^2 based association	$\sum_{v \in (w_1, \bar{w}_1)} \frac{(f(vu) - f_\emptyset(vu))^2}{f(vu)}$

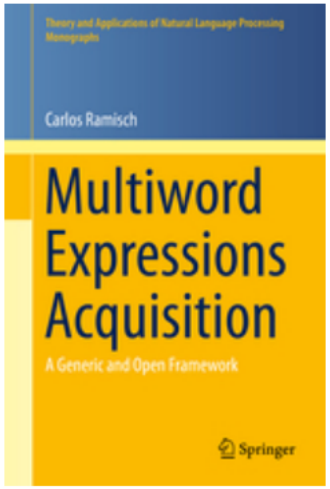
Collocational preferences

- Generate list of candidate MWEs from a corpus
  - n-grams (Manning and Schütze 1999)
  - syntactic patterns (Justeson and Katz 1995)
- Rank candidates by score of association strength,
  - stronger associations expected to be genuine MWEs
- Combine with other sources of information
  - Syntactic analysis (Seretan 2011)
  - Translations (Caseli et al. 2010, Attia et al. 2010, Tsvetkov and Wintner 2010)

Collocational preferences

mwetoolkit^3

http://mwetoolkit.sourceforge.net/PHITE.php



Home

- Install
  - Linux
  - Mac OS
  - Windows
- Tutorials
- Tools
- File types
- For developers
- Contact

Installing the `mwetoolkit`

1. Checking the requirements

Before following the instructions below, please check that you have installed all the requirements according t your Operating System:

- Linux
- Mac OS
- Windows

Once the requirements are installed, choose one installation method, from GIT or from release, and follow th instructions below. As the code evolves fast and releases are not that frequent, we HIGHLY recommend you t use the GIT version.

VPCs in Child Language

- English CHILDES corpora (MacWhinney, 1995)
- Verb-particle constructions (VPCs) identified from verbs separated from particles by up to 5 words (Baldwin, 2005)

Sentences	Children Set	Adults Set
Parsed	482,137	988,101
with VPCs	44,305	83,098
with VPCs Cleaned	38,326	82,796

Age in months	VPC Sentences
0-24	2,799
24-48	26,152
48-72	8,038
72-96	1,337
>96	514
No age	4,841

Get out but don't fall down: verb-particle constructions in child language

Aline Villavicencio[♣♠], Marco A. P. Idiart[♣], Carlos Ramisch[♣], Vítor Araújo[♣], , Beracah Yankama[♠], Robert Berwick[♠]

[♣]Federal University of Rio Grande do Sul (Brazil)
[♠]MIT (USA)
alinev@gmail.com, marco.idiart@gmail.com, ceramisch@inf.ufrgs.br, vbuaraujo@inf.ufrgs.br, beracah@mit.edu, berwick@csail.mit.edu

Abstract

Much has been discussed about the challenges posed by Multiword Expressions (MWEs) given their idiosyncratic, flexible and heterogeneous nature. Nonetheless, children successfully learn to use them and eventually acquire a number of Multiword Expressions comparable to that of

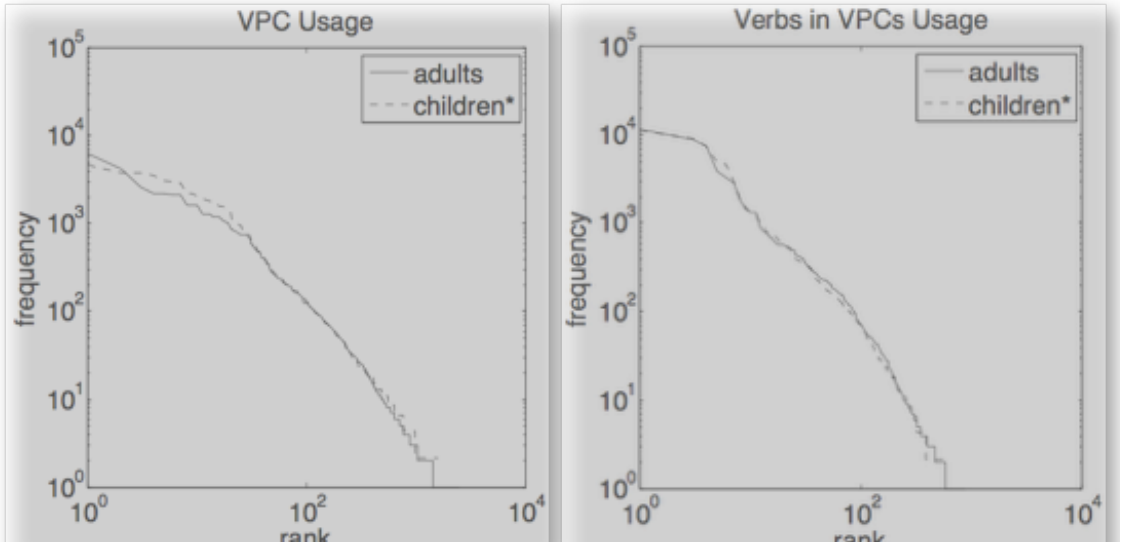
such as terminology and formulaic linguistic units (Wray, 2009). Depending on the definition, they may also include less traditional sequences like *copy of* in *They gave me a copy of the book* (Fillmore et al., 1988), greeting formulae like *how do you do?*, and lexical bundles such as *I dont know whether* or memorized poems and familiar phrases from TV commercials (Jackendoff

Aline Villavicencio, Marco Idiart, Carlos Ramisch, Vitor Araujo, Beracah Yankama, Robert Berwick, "Get out but don't fall down: verb-particle constructions in child language", Proceedings of the Workshop on Computational Models of Language Acquisition and Loss, Avignon, France, 2012.

VPCs in Child Language

- Similar production rates
  - 7.95% (children) vs. 8.38% (adults)
- Similar frequencies per bin
  - Zipfian distribution
    - adult rank = children rank * 2.16 between VPC tokens by adults and children

Frequency	Children Set	Adults Set
1	42.62%	43.03%
2	13.05%	15%
3	8.36%	6.48%
4	4.05%	4.5%
≥5	31.92%	31%



VPCs in Child Language

- Children vs. Adult
  - VPCs types: Kendall τ score = 0.63
  - Verbs in VPCs: Kendall τ score = 0.84

Rank	Chidren VPC	Children Freq	Adult VPC	Adult Freq	Child Rank
1	put on	2005	come on	6244	7
2	go in	1608	put on	4217	1
3	get out	1542	go on	2660	9
4	take off	1525	get out	2251	3
5	fall down	1329	take off	2249	4
6	put in	1284	put in	2177	6
7	come on	1001	sit down	2133	8
8	sit down	981	go in	1661	2
9	go on	933	come out	1654	10
10	come out	872	pick up	1650	18

Top 10 VPCs

  - Distance: over 97% of VPCs have at most intervening 2 words

Distance	Children Set	Adults Set
0	65.13%	64.14%
1	23.48%	22.15%
2	9.33%	10.90%
3	1.65%	2.15%
4	0.29%	0.47%
5	0.09%	0.16%

CANONICAL FORM PREFERENCES

Canonical Form Preferences

- MWEs have greater fixedness in comparison with ordinary word combinations (Sag et al. 2002)
  - *to make ends meet* (to earn just enough money to live on)
    - Choice of determiner:
      - *?to make some/these/many ends meet*
    - Pronominalisation:
      - *?make them meet*
    - Internal modification:
      - *?to make ends quickly meet*

Canonical Form Preferences

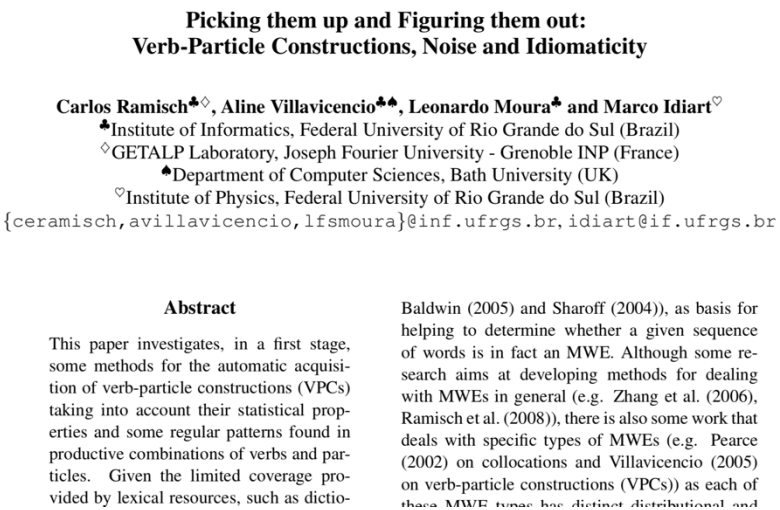
- Fixedness detection:
  - Generate expected variants and compare with observed variants
    - Limited degree of variation for idiomatic MWEs (Ramisch et al. 2008, Geeraert et al. 2017)
    - Preference for canonical form for idiomatic MWEs (Fazly et al. 2009, King and Cook 2018)
    - Less similarity with variants for idiomatic MWEs in DSMs (Senaldi et al. 2019)
  - Lexical substitution variants:
    - WordNet (Pearce 2001; Ramisch et al. 2008, Senaldi et al.2019)
    - Levin's semantic classes (Villavicencio 2005; Ramisch et al. 2008)
    - Distributional Semantic Models (Senaldi et al. 2019)

VPC Discovery

- Entropy-based measure of canonical form preference
  - Compositional VPCs have more variants (high entropy)
    - VPC: Precision: 0.85, Recall: 0.96, F-measure: 0.90
    - Idiomaticity: Precision: 0.62, Recall: 0.25

$$H(V) = -\sum^{n} p(v_i) \ln [\, p(v_i) \,]$$

$$p(v_i) = \frac{n(v_i)}{\sum_{\forall\, v_j \in V} n(v_j)}$$



Picking them up and Figuring them out:
Verb-Particle Constructions, Noise and Idiomaticity

Carlos Ramisch[♣♢], **Aline Villavicencio**[♣♠], **Leonardo Moura**[♣] **and Marco Idiart**[♡]
[♣]Institute of Informatics, Federal University of Rio Grande do Sul (Brazil)
[♢]GETALP Laboratory, Joseph Fourier University - Grenoble INP (France)
[♠]Department of Computer Sciences, Bath University (UK)
[♡]Institute of Physics, Federal University of Rio Grande do Sul (Brazil)
{ceramisch,avillavicencio,lfsmoura}@inf.ufrgs.br, idiart@if.ufrgs.br

Abstract

This paper investigates, in a first stage, some methods for the automatic acquisition of verb-particle constructions (VPCs) taking into account their statistical properties and some regular patterns found in productive combinations of verbs and particles. Given the limited coverage provided by lexical resources, such as dictio-

Baldwin (2005) and Sharoff (2004)), as basis for helping to determine whether a given sequence of words is in fact an MWE. Although some research aims at developing methods for dealing with MWEs in general (e.g. Zhang et al. (2006), Ramisch et al. (2008)), there is also some work that deals with specific types of MWEs (e.g. Pearce (2002) on collocations and Villavicencio (2005) on verb-particle constructions (VPCs)) as each of

Carlos Ramisch, Aline Villavicencio, Leonardo Moura, Marco Idiart, "**Picking them up and Figuring them out: Verb-Particle Constructions, Noise and Idiomaticity**"*CoNLL 2008*, Manchester, UK, 2008.

In this talk

- Collocational Properties
- Canonical Form Preferences
- Contextual Preferences
- Conclusions and Future Work

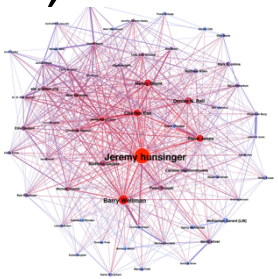
CONTEXTUAL PREFERENCES

Contextual Preference

- You shall know a (multi)word by the company it keeps (adaption of Firth 1957)
  - Assumptions
    1. Words can be characterised by contexts
       - *Famous author **writes** book under a pseudonym*
       - we can approximate MWE meaning by compiling affinities with contexts
    2. Words that occur in similar contexts have similar meanings (Turney and Pantel 2010)
       - *author **writes/rewrites/composes/creates/prepares** book*
       - we can find (multi)words with similar meanings measuring how similar their contextual affinities are

Contextual preferences

- Distributional semantic models (or vector space models)
  - Represent meaning as numerical multidimensional vectors in semantic space
    - Lin 1998; Pennington et al. 2014; Mikolov et al. 2013, Peters et al 2018, Joshi et al. 2019
  - Reach high levels of agreement with human judgments about word similarity
    - Baroni et al. 2014; Camacho-Collados et al. 2015; Lapesa and Evert 2017

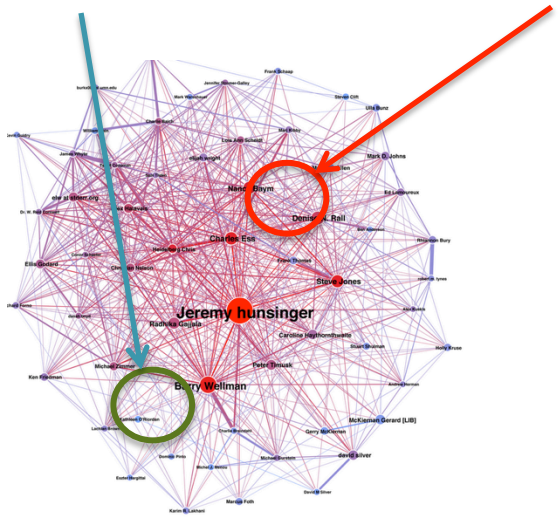


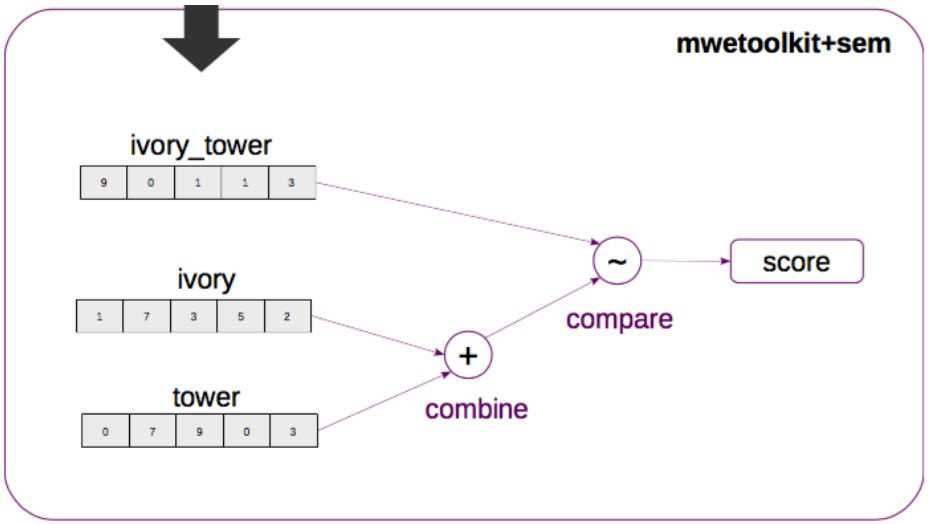
Contextual preferences

- DSMs use algebra to model complex interactions between words
  - Vectors of MWE components composed
    - Additive model (Mitchell and Lapata 2008)
      - Parameters for importance of meaning of part (Reddy et al. 2011)
        - » *flea market*: head (*market*) contributes more to meaning
    - Other operations (Mitchell and Lapata 2010; Reddy et al. 2011; Mikolov et al. 2013; Salehi et al. 2015; Cordeiro et al. 2019)
  - Similarity or relatedness modelled as comparison between word vectors

Contextual preferences

- Cosine similarity between the MWE vector and the sum of the vectors of the component words
  - **cos($w_1w_2$vector, $w_1$vector+$w_2$vector)**





- Distance indicates degree of idiomaticity
  - the closer they are, the more compositional the MWE

How to detect compositionality?

- To what extent the meaning of MWE can be computed from the meanings of component words using DSMs
  - Is accuracy in prediction dependent on
    - characteristics of the DSMs ?
    - the language/corpora ?

How to detect compositionality?

- Over 9,000 analyses and 680 DSMs detailed in

Silvio Cordeiro, Aline Villavicencio, Marco Idiart, Carlos Ramisch, "**Unsupervised Compositionality Prediction of Nominal Compounds**", *Computational Linguistics*, 45(1):1--57, 2019, MIT Press.

Unsupervised Compositionality Prediction of Nominal Compounds

Silvio Cordeiro*
Federal University of Rio Grande do Sul and Aix Marseille Univ, CNRS, LIS

Aline Villavicencio**†
University of Essex and Federal University of Rio Grande do Sul

Marco Idiart‡
Federal University of Rio Grande do Sul

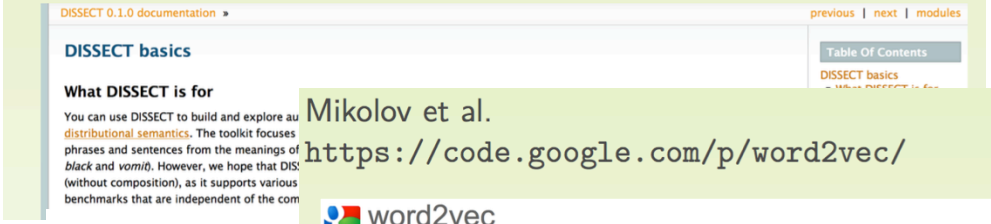
Carlos Ramisch§
Aix Marseille Univ, CNRS, LIS

Distributional Semantic Models

- Constructing DSMs
  - Dissect (Dinu et al., 2013), Minimantics (Ramisch et al. 2013), word2vec (Mikolov et al., 2013) and Glove (Pennington et al., 2014).

dissect

Baroni et al. http://clic.cimec.unitn.it/composes/toolkit/introduction.html



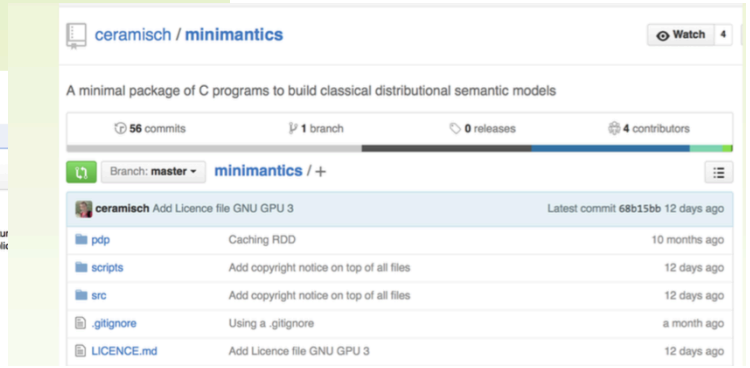
GloVe

Mikolov et al.
https://code.google.com/p/word2vec/

word2vec

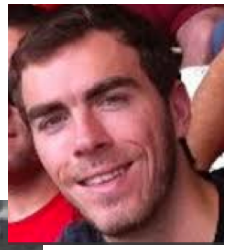
Minimantics

https://github.com/ceramisch/minimantics



Distributional Semantic Models

- LexVec (Lexical Vectors)
  - Alternative that word2vec and GloVe in word similarity tasks
    - Freely available



ACL 2016

Matrix Factorization using Window Sampling and Negative Sampling for Improved Word Representations

Alexandre Salle[1] **Marco Idiart**[2] **Aline Villavicencio**[1]
[1] Institute of Informatics
[2] Physics Department
Universidade Federal do Rio Grande do Sul
Porto Alegre, Brazil
{atsalle,avillavicencio}@inf.ufrgs.br, idiart@if.ufrgs.br

Abstract

In this paper, we propose LexVec, a new method for generating distributed word representations that uses low-rank, weighted factorization of the Positive Point-wise Mutual Information matrix via

In this paper, we present Lexical Vectors (LexVec), a method for factorizing PPMI matrices that combines characteristics of all these methods. On the one hand, it uses SGNS window sampling, negative sampling, and stochastic gradient descent (SGD) to minimize a loss function that weights frequent co-occurrences heavily but also

The models

- DSMs
  - PPMI models – positive PMI (Minimantics)
  - GloVe (Pennington et al. 2014)
  - Word2vec (Mikolov et al 2013) Skipgram, CBOW
  - LexVec (Salle et al. 2016, 2018)
- WaCky Corpora (Baroni et al., 2009):
  - ukWaC for English (~2 billion tokens)
  - frWaC (~1.6 billion tokens) for French
  - brWaC (~2.3 billion tokens) for Portuguese (Wagner Filho et al. 2016)
- Pre-processing
    - *surface+*: the original corpus
    - *surface*: with stopword removal.
    - *lemma*: stopword removal and lemmatization;
    - *lemmaPOS*: stopword removal, lemmatization and POS-tagging
- Context Window size: 1,4 and 8
- Dimension size: 250, 500, 750

Gold Standards

- Roller et al. (2013) 244 German compounds
  - around 30 judgments by crowdsourcing
  - scale from 1 to 7
- Farahmand et al. (2015) 1,042 English compounds
  - 4 experts judges
  - binary scale for non-compositionality and conventionality
- Reddy et al. (2011) 90 English compounds
  - around 30 judgments by crowdsourcing
  - scale from 0 to 5
- Kruszewski and Baroni (2014) 5,849 judgments for modifier-head phrases in English
  - Is the phrase an instance of concept denoted by head (*dead parrot* and *parrot*)
  - Is it a member of more general concept that includes head (*dead parrot* and *pet*),
  - typicality ratings,
- We used Reddy's protocol as basis to add 180 compounds and expand to other languages

Collecting Human Judgments

- Multilingual dataset with 180 compounds in each language
  - English: N_1 N_2
    - *olive oil*
    - extends Reddy et al. 2011 with 90 compounds
  - French: N_2 A_1
    - *mort cellulaire* (*cell death)*
  - Portuguese: N_2 A_1
    - *morte celular* (*cell death)*
- Balanced for compositionality
  - 60 idiomatic, 60 partially compositional and 60 compositional

ACL 2016

How Naked is the Naked Truth?
A Multilingual Lexicon of Nominal Compound Compositionality

Carlos Ramisch[1], **Silvio Cordeiro**[1,2], **Leonardo Zilio**[2]
Marco Idiart[3], **Aline Villavicencio**[2], **Rodrigo Wilkens**[2]
[1] Aix Marseille Université, CNRS, LIF UMR 7279 (France)
[2] Institute of Informatics, Federal University of Rio Grande do Sul (Brazil)
[3] Institute of Physics, Federal University of Rio Grande do Sul (Brazil)
silvioricardoc@gmail.com carlos.ramisch@lif.univ-mrs.fr lzilio@inf.ufrgs.br
marco.idiart@gmail.com avillavicencio@inf.ufrgs.br rswilkens@inf.ufrgs.br

Project FAPERGS-CNRS-INRIA (France Brazil)

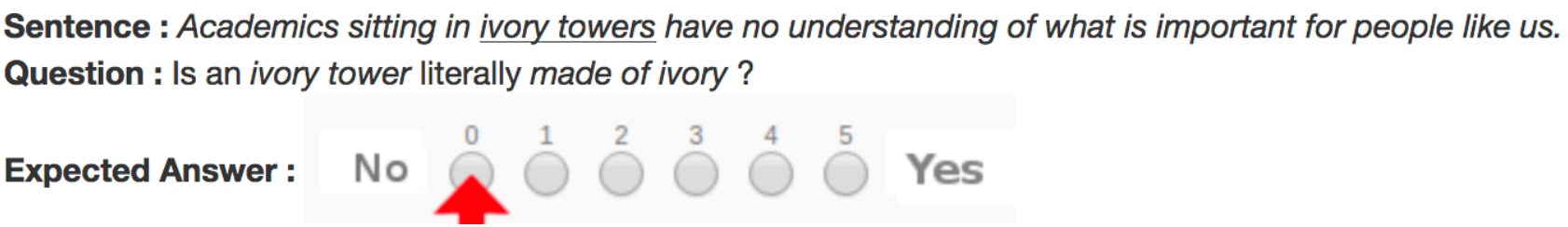
Collecting Human Judgments

- Following Reddy et al. (2011) use literality to approximate compositionality
- Judgments with likert scale (0 to 5)
  - For **compound**

Sentence : *Policies designed to encourage adaptation to climate change may conflict with regulation aimed at protecting the environment.*
Question : Is *climate change* truly/literally a *change* in *climate* ?

Collecting Human Judgments

- Following Reddy et al. (2011) use literality to approximate compositionality
- Judgments with likert scale (0 to 5)
  - For compound
  - For $\mathbf{w_1}$ and for w_2 separately



Collecting Human Judgments

- Context: 3 sentences per compound
  - Compound has same meaning in all sentences
- Participants: linguists, CS students, AMT workers

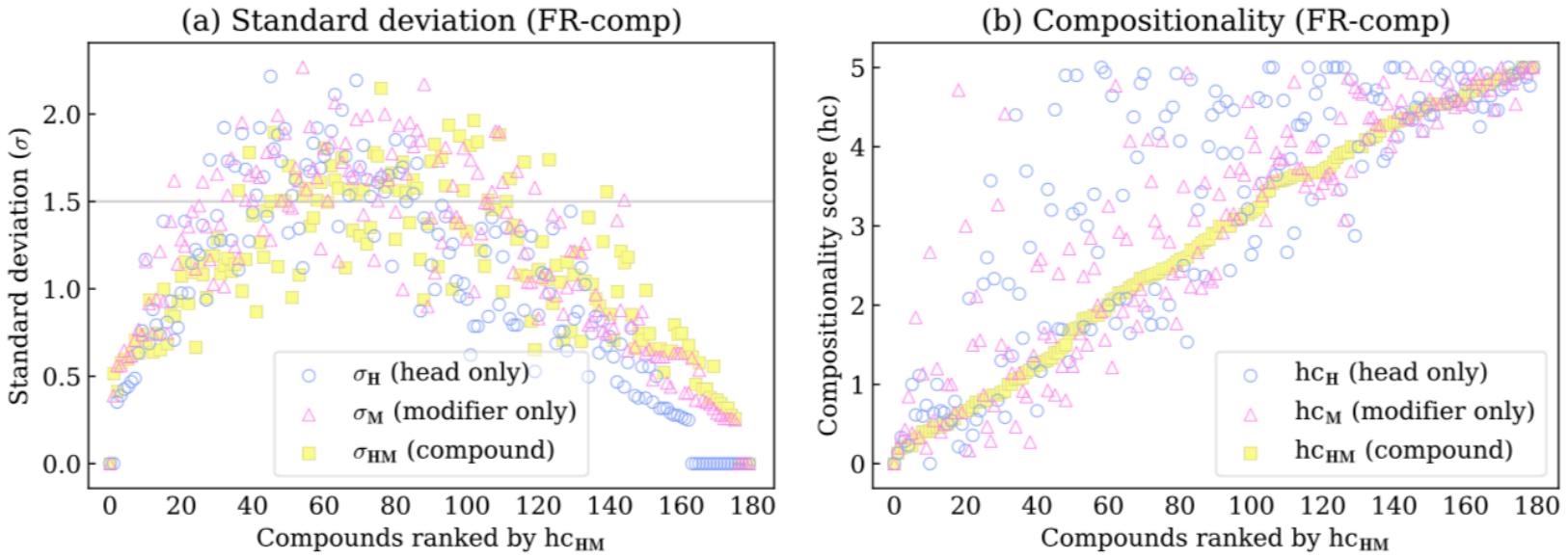
Collecting Human Judgments - Agreement

- For Portuguese subset of annotators
  - α = .52 for head,
  - α = .36 for modifier
  - α = .42 for compound
  - Same annotator after 1 month:
    - α = .59 for compound
    - ρ = .77 for compound
      - qualitative upper bound for compositionality prediction on *PT-comp*.
- Average standard deviation in judgments

Data set	$\overline{n}$	$\overline{\sigma_{\text{HM}}}$
FR-comp	14.9	1.15
PT-comp	31.8	1.22
EN-comp$_{90}$	18.8	1.17
EN-comp$_{Ext}$	22.6	1.21
Reddy	28.4	0.99

Collecting Human Judgments - Agreement

- Greater agreement between score for compound and head (or modifier) for extremes
  - totally idiomatic and fully compositional
- Asymmetric impact of non-literal part: score determined by the least literal word



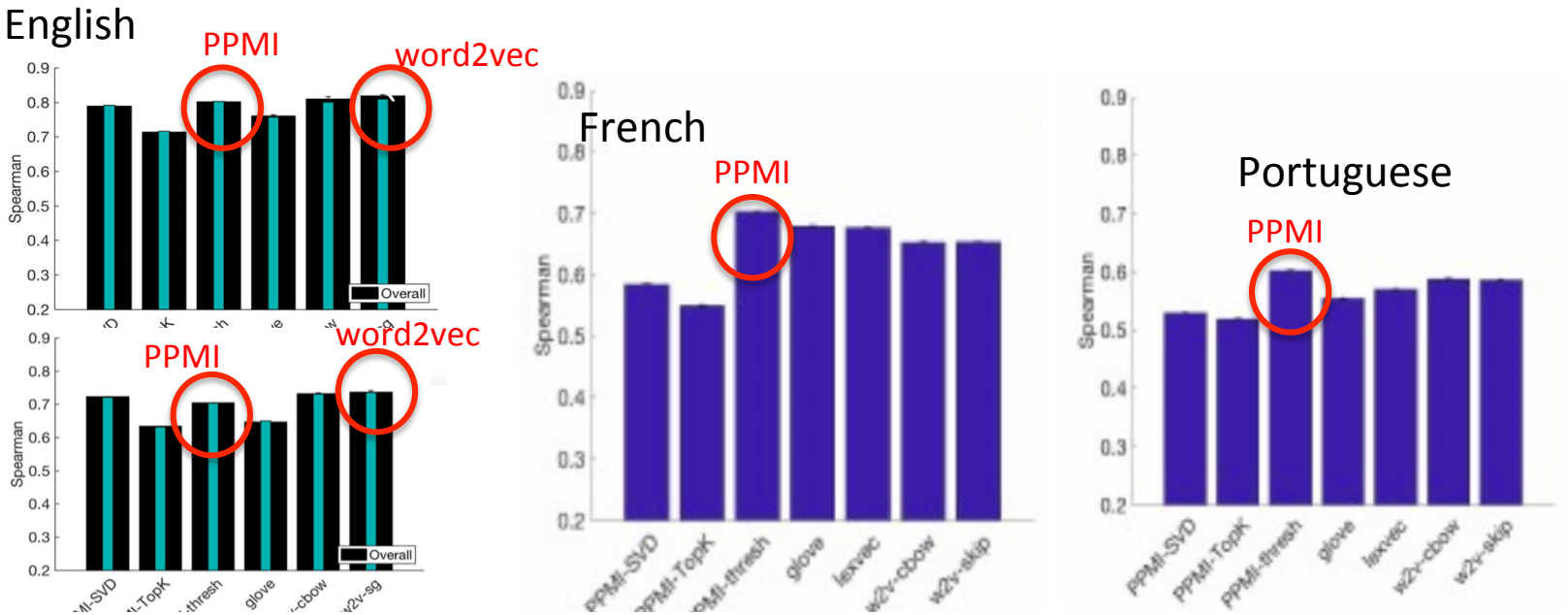
Agreement

- Most/least variation in scores (average±σ score)

	compound	head	mod	comp
English	brass ring	3.9 ±2.0	3.7 ±1.9	3.7 ±1.8
	fish story	4.8 ±0.4	1.5 ±1.8	1.7 ±1.8
	tennis elbow	4.3 ±1.3	2.2 ±1.8	2.5 ±1.8
	brick wall	3.5 ±1.9	3.2 ±2.2	3.8 ±1.7
	dirty word	4.1 ±1.4	2.0 ±1.4	2.5 ±1.7
	prison guard	4.8 ±0.4	4.9 ±0.3	4.9 ±0.3
	graduate student	5.0 ±0.0	4.7 ±0.5	4.9 ±0.3
	engine room	5.0 ±0.0	4.9 ±0.3	4.9 ±0.3
	climate change	4.8 ±0.4	4.9 ±0.3	5.0 ±0.2
	insurance company	4.9 ±0.5	5.0 ±0.0	5.0 ±0.0

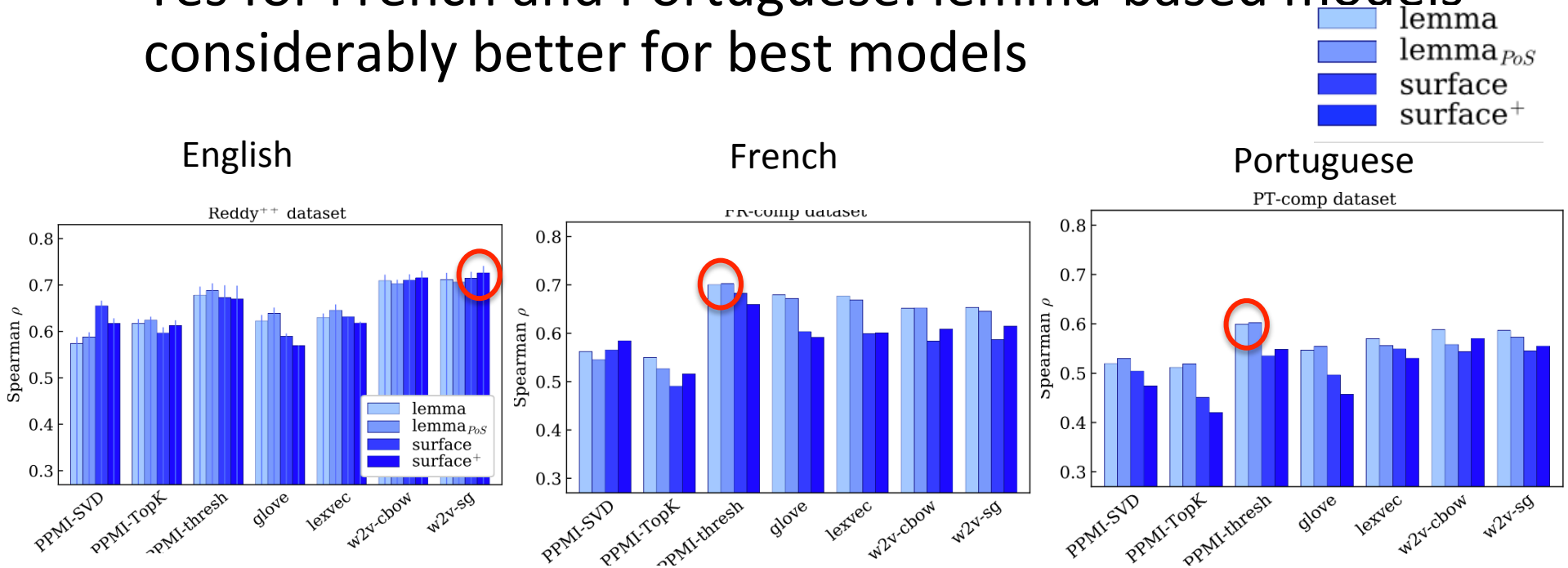
Evaluation

- Comparing model predictions with average human judgment
  - English Reddy: word2vec, Spearman ρ=0.82
  - English Reddy++: word2vec, Spearman ρ=0.73
  - French: PPMI global context, Spearman ρ=0.70
  - Portuguese: PPMI global context, Spearman ρ=0.60



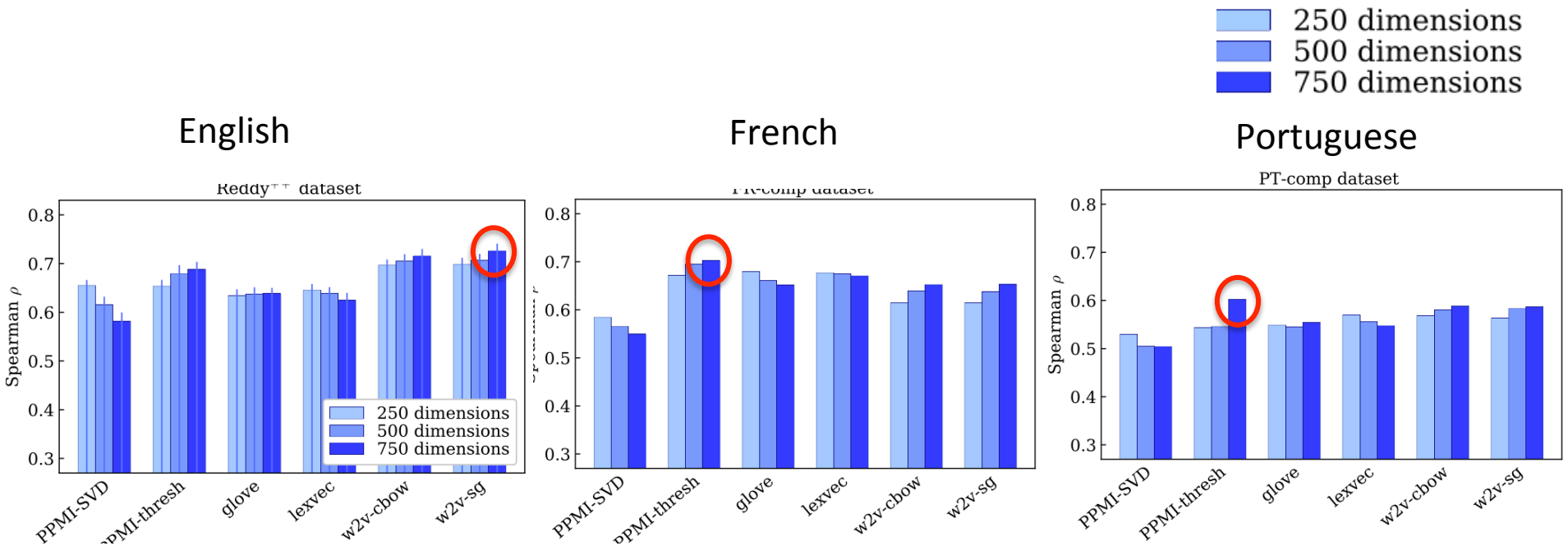
Evaluation − Type of Preprocessing

- Do less sparse representations lead to better results?
  - Not for English: preprocessing makes no differences for best model
  - Yes for French and Portuguese: lemma-based models considerably better for best models



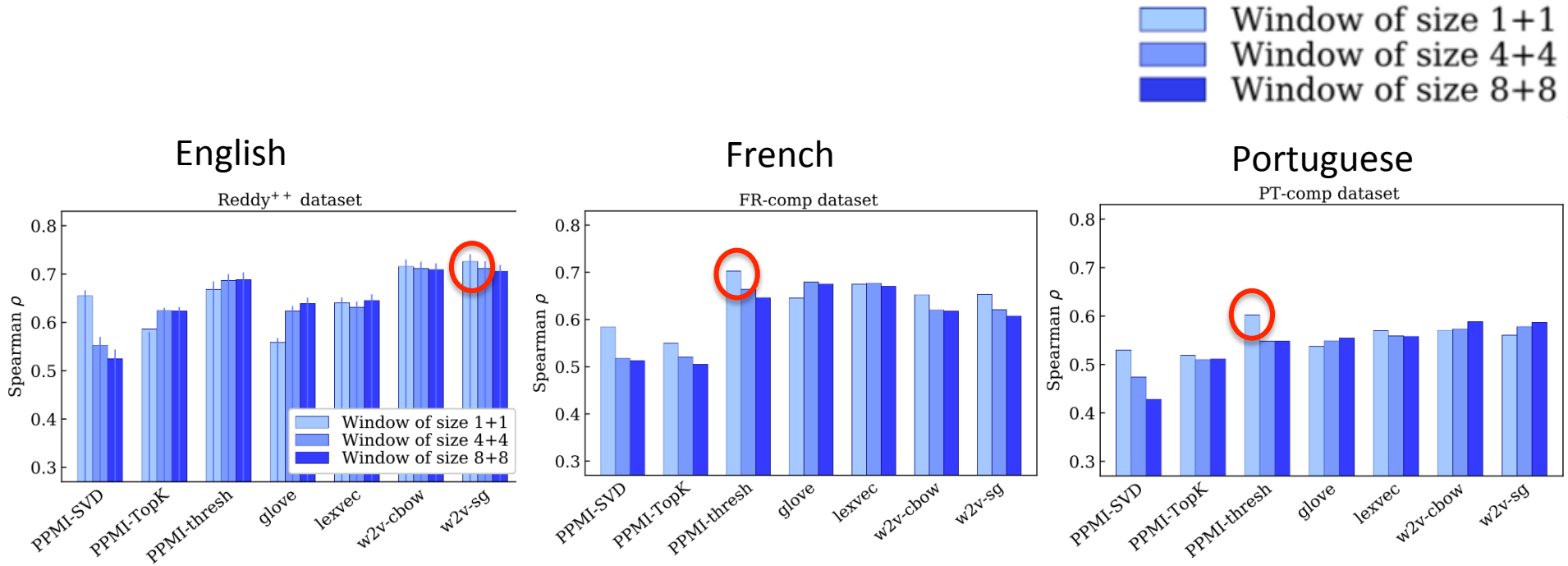
Evaluation – Number of Dimensions

- Do larger dimensions lead to more accurate models/better results?
  - Yes for English, French and Portuguese: more dimensions lead to better results

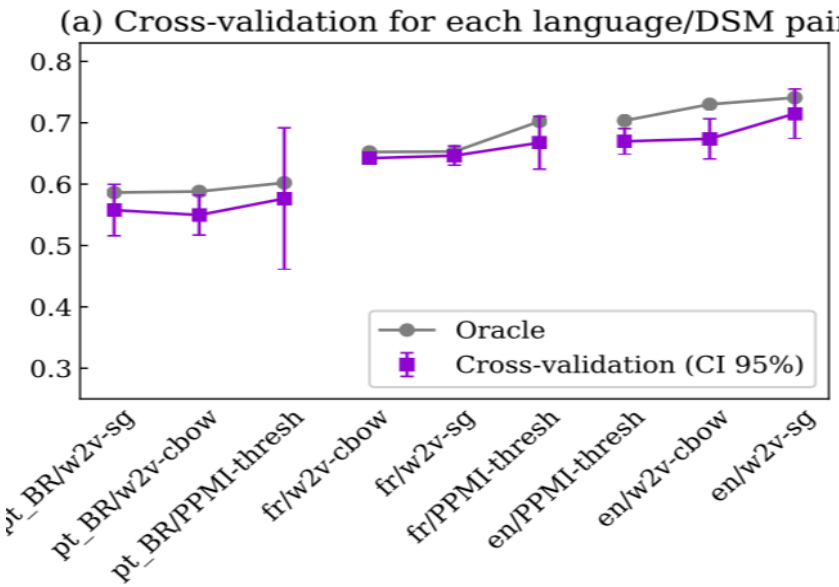


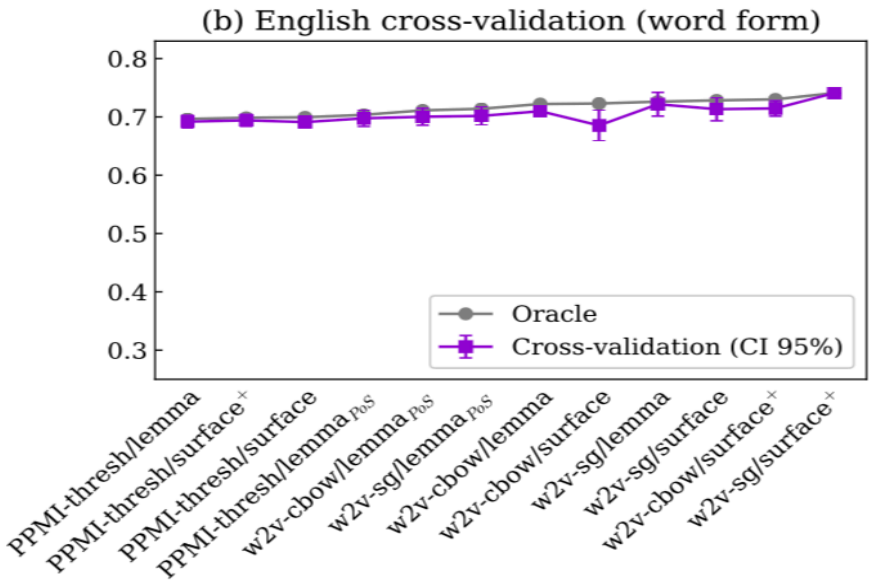
Evaluation − Size of Context Window

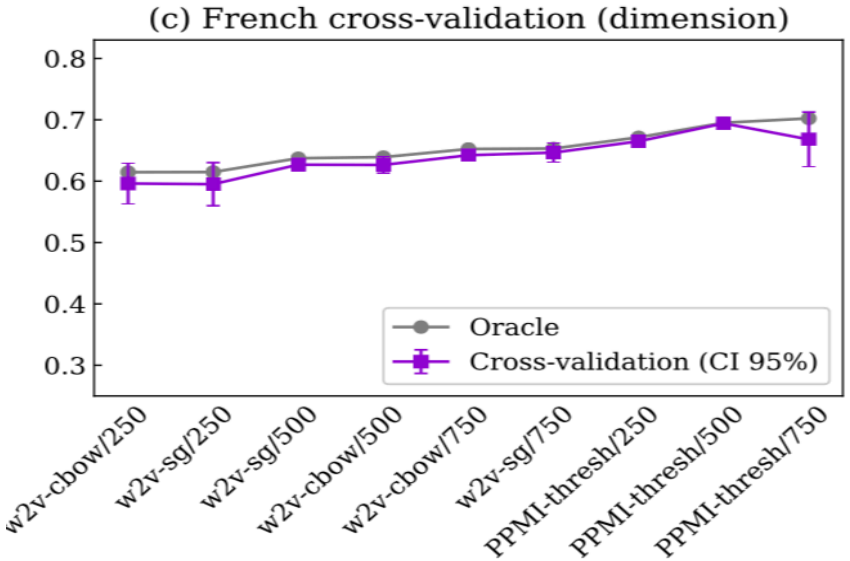
- Do larger window sizes lead to better results?
  - Not for English, French and Portuguese: trend for smaller windows in best models

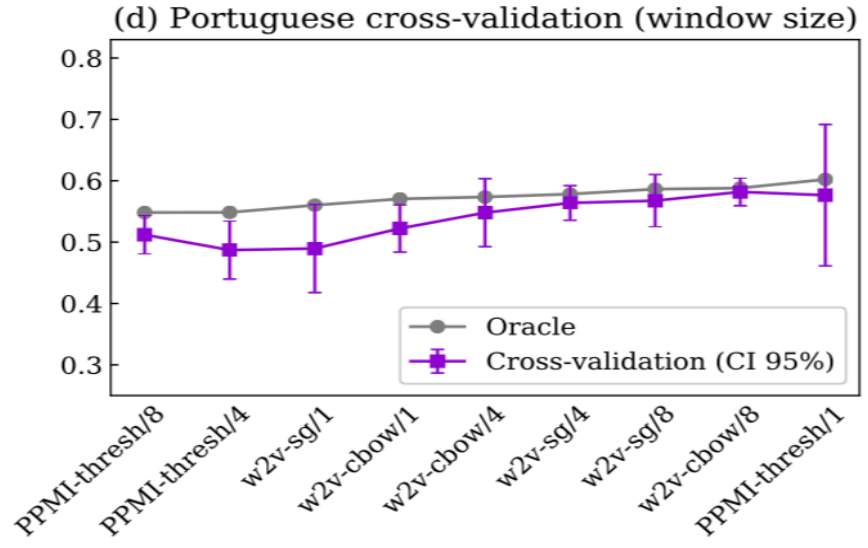


Evaluation - Cross-validation



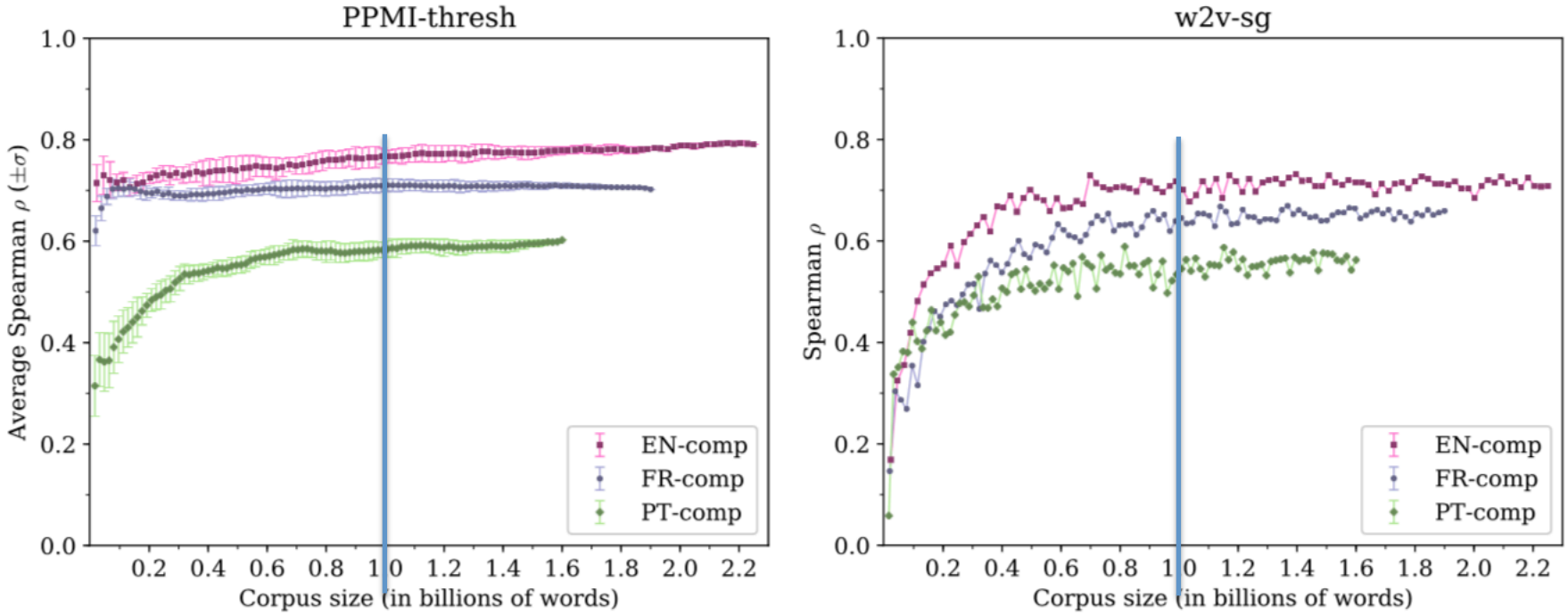






Evaluation − Corpus Size

- Are better results for English due to larger corpus size?



  - Not for English, French and Portuguese:
    - stable performance after ~1 billion words
      - all compounds may be frequent enough for accurate representations

CONCLUSIONS

DSMs and Compositionality

- Large-scale multilingual analysis of DSMs for compound compositionality prediction
  - in English, French and Portuguese
  - Over 600 DSMs and
  - Almost 9000 evaluations
  - 3 families of models: word2vec, GloVe, and PPMI-based models.

DSMs and Compositionality

- Dataset of nominal compounds with human judgments about literality/compositionality
  - 270 compounds for English,
  - 180 for French and Portuguese

Compositionality of Nominal Compounds - Datasets

- Authors: Silvio Cordeiro, Carlos Ramisch, Aline Villavicencio, Leonardo Zilio, Marco Idiart, Rodrigo Wilkens
- Version 1.0 - August 2, 2016
- Download the data set

Description

This package contains numerical judgements by human native speakers about 180 nominal compound compositionality in English (EN), French (FR) and Brazilian Po

Judgements were obtained using Amazon Mechanical Turk (EN and FR) and a web interface for volunteers (PT). Every compound has 3 scores: compositionality of h (fully compositonal) and are averaged over several annotators (around 10 to 20 depending on the language). All compounds in FR and PT, and 90 compounds in EN,

The datasets are described in detail and used in the experiments of papers below. Please cite one of them if you use this material in your research.

- How Naked is the Naked Truth? A Multilingual Lexicon of Nominal Compound Compositionality [bib]
- Predicting the Compositionality of Nominal Compounds: Giving Word Embeddings a Hard Time [bib]
- Filtering and Measuring the Intrinsic Quality of Human Compositionality Judgments [bib]

Our methodology is inspired from Reddy, McCarthy and Manandhar (2011). We include their set of 90 compounds and judgments in our datase full EN dataset.

Quick start

If you only want to use our datasets to evaluate your compositionality prediction models, you're probably interested in the scores present in t itionality of files:

- **annotations/en.unfiltered.csv**
- **annotations/fr.unfiltered.csv**
- **annotations/pt.unfiltered.csv**

DSMs and Compositionality

- Dataset of Lexical Substitution of Nominal Compounds in Portuguese (LexSubNC)
  - 180 compounds for Portuguese
  - Resource freely available

LexSubNC - Lexical Substitution of Nominal Compounds in Portuguese

- *Rodrigo Wilkens, Leonardo Zilio, Silvio Cordeiro, Felipe S. F. Paula, Carlos Ramisch, Marco Idiart, Aline Villavicencio*
- *Version 1.0 - September 20, 2017*
- Download the data set

Description

This package is an extension of the original compositionality datasets and includes more detailed annotation for Portuguese lexical substitution candidates in the original data
compounds in Portuguese as the compositionality dataset. It additionally contains frequency and PMI from a large Brazilian Portuguese corpos (around 1.2 billion words), as w
the following categories:

- Invalid: the substitution candidate is not fit for substitution, either for being too specific for a given context or for simply not being valid for the target MWE.
- Syn-SW: the substitution candidate is a single-word matching synonym in relation to the target MWE.
- NearSyn-SW: the substitution candidate is a single-word quasi-synonym in relation to the target MWE.
- Syn-MWE: the substitution candidate is a multiword matching synonym in relation to the target MWE.
- NearSyn-MWE: the substitution candidate is a multiword quasi-synonym in relation to the target MWE.
- Paraphrase: the substitution candidate is a paraphrasis of the target MWE.
- Definition: the substitution candidate is a definition of the target MWE.
- Head
- Modifier

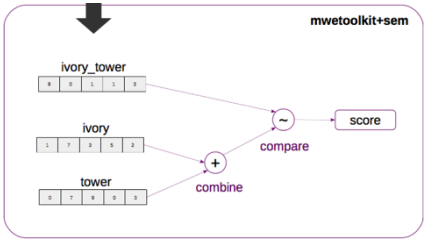
LexSubNC: A Dataset of Lexical Substitution for Nominal Compounds

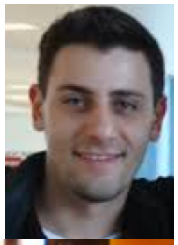
- Noun compound substitutes collected through **crowdsourcing**
  - 180 Portuguese compounds
  - 3,061 substitutes in context

Classification	# Total Responses	# Unique Responses	Compound Coverage
Syn_{word}	966	318	99/180
Syn_{MWE}	1,257	684	159/180
$NearSyn_{word}$	315	150	83/180
$NearSyn_{MWE}$	303	183	96/180
Paraphrase	54	47	24/180
Definition	166	162	90/180
Total	3,061	1,544	180/180

mwetoolkit

- Language independent framework for MWE processing
- Extracts MWE from corpora
- Annotates corpora with MWEs
- Calculates AMs
- Pre-processes MWEs in corpora for DSM construction
- Imports DSMs (word2vec, glove, PPMI)
- Provides functions for vector combinations
- Calculates compositionality
- Evaluates against gold standard





LREC 2016

mwetoolkit+sem: Integrating Word Embeddings in the mwetoolkit for Semantic MWE Processing

Silvio Cordeiro[1,2], **Carlos Ramisch**[2], **Aline Villavicencio**[1]
[1] Institute of Informatics, Federal University of Rio Grande do Sul (Brazil)
[2] Aix Marseille Université, CNRS, LIF UMR 7279 (France)
silvioricardoc@gmail.com carlos.ramisch@lif.univ-mrs.fr avillavicencio@inf.ufrgs.br

Abstract

This paper presents mwetoolkit+sem: an extension of the mwetoolkit that estimates semantic compositionality scores for multiword expressions (MWEs) based on word embeddings. First, we describe our implementation of vector-space operations working on distributional vectors. The compositionality score is based on the cosine distance between the MWE vector and the composition of the vectors of its member words. Our generic system can handle several types of word embeddings and MWE lists, and may combine individual word representations using several composition techniques. We evaluate our implementation on a dataset of 1042 English noun compounds (Farahmand et al., 2015), comparing different configurations of the underlying word embeddings and

CNPq CAPES

Project CAPES-COFECUB (France-Brazil)

Future Work

- More accurate (multi)word representations
  - ACL 2019: Jana et al. 2019, Qi et al. 2019
- Token idiomaticity identification
  - Gharbieh et al. 2017, Taslimipoor et al.l 2017, King and Cook 2018,
- Machine Translation
  - Kick the bucket → morrer/*chutar o balde

This research was done in collaboration with Carlos Ramisch, Marco Idiart, Silvio Cordeiro, Rodrigo Wilkens and Leonardo Zilio

This work was partly supported by the Brazilian Research Council (CNPq 423843/2016-8) and by the Human Rights, Big Data and Technology Project (University of Essex).

THANK YOU

When the whole is greater than the sum of its parts: Multiword expressions and idiomaticity

Aline Villavicencio

University of Essex (UK)

Federal University of Rio Grande do Sul (Brazil)