Predicting the Semantic Compositionality of Prefix Verbs

Shane Bergsma, Aditya Bhargava, Hua He and Grzegorz Kondrak

University of Alberta

EMNLP 2010

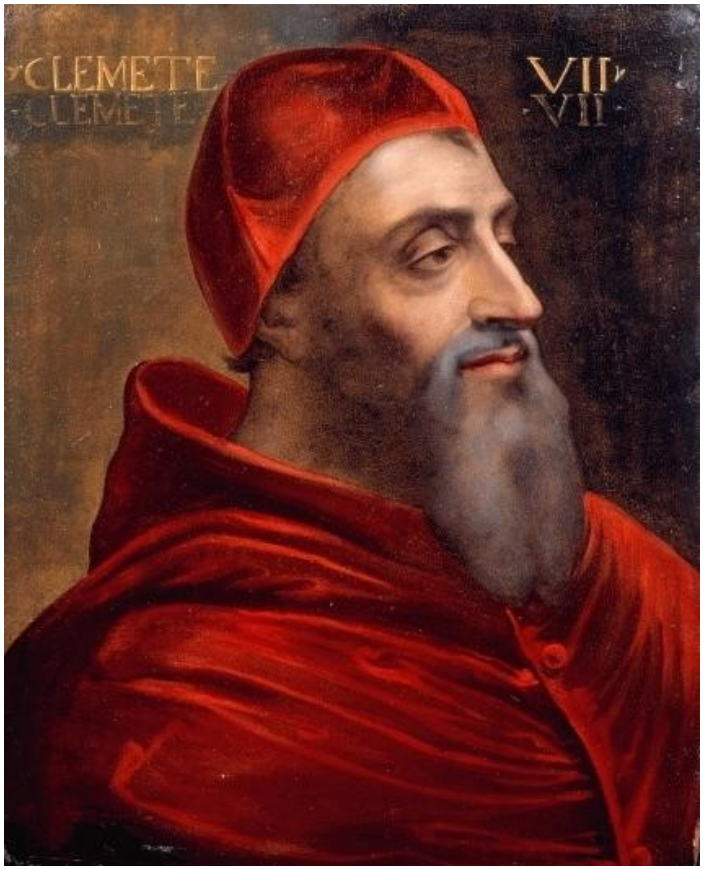
Prefix Verbs

- **Definition**:
  - ***prefix*** (***re-,out-,under-***) + ***verb-stem***
  - e.g. ***re*marry**, ***re*tire**, ***out*swim**, ***under*stand**
- **Objective**: determine whether the meaning of a prefix verb is compositional
- **Method**: supervised classifier with a range of interesting distributional features
- **Results**: **94% accuracy**

Motivation

- Rephrasing complex forms with stems reduces sparsity and reveals connections
- Our goal: Adopt and apply a definition of compositionality that supports downstream applications
  - E.g. information retrieval (IR), textual entailment, text categorization, statistical machine translation, etc.

Question Answering Example



Corpus:

... Pope Clement VII denied Henry VIII permission to **marry** again...

Question Answering Example

Corpus:

... Pope Clement VII denied Henry VIII permission to **marry** again...

Query:

Which Pope refused Henry VIII permission to ***re*marry**?

Previous Work

- Morphology: break word into *morphemes*: smallest meaning-bearing units
  - compositionality is part of the story
- Stemming, lemmatization, compound splitting (e.g. in German) for IR
- Porter stemmer [Porter '80], PC-KIMMO [Karp et al. '92], *morpha* [Minnen et al. '01], etc., only handle word *suffixes*

Applications of Morphology

- "Full morphological analysis provides at most very modest benefits for retrieval."

  [Manning et al., 2008]

- Most morpho work in NLP not designed for applications, but aims to match output of a *human morphologist*
- But ***cooperate*** is not exactly ***co*+operate**:
  - E.g. "Which nations ***cooperate*** with the ICC?"

Our Definition

- Prefix-verb compositionality is a semantic equivalence between a prefix verb and a paraphrase involving the verb's stem
  - e.g. ***out*build ↔ build more/better/faster** than
- Compositionality is context-dependent
  - a property of tokens in a given sentence
  - E.g. "***re*sort** to force", but "***re*sort** a linked list"
- Type-based solution works okay

Supervised Classification

- Train classifier on data annotated according to our definition (use SVMs)
- Key contribution: features derived from web-scale N-gram data
- Leverage semantic as well as orthographic information [Yarowsky & Wicentowski ‘00, Schone & Jurafsky ‘01, Baroni et al. ’02]
- Similar work for MWEs [Baldwin et al. ’03]

Features: Hyphenation Count

- Compositional prefix verbs are often hyphenated
- Is the verb "***re*elect**" compositional?

  frequency("re-elect") = **33,000**

  frequency("reelect") = **9,000**

  Yes
- Is the verb "***re*tire**" compositional?

  frequency("re-tire") = **121**

  frequency("retire") = **1,115,000**

  No

Features: Co-occurrence

- Compositional prefix verbs often co-occur with their separated stems
  - "**elect**" and "**reelect**" co-occur in 314 N-grams
    - "The voters, who **elect** and **reelect** district attorneys or chief prosecutors..."
  - "**tire**" and "**retire**" occur in only 12 N-grams
    - "Mom would **tire** and ***re*tire** to the kitchen."
- Via N-grams and *Yahoo Search API*

Other Features

- Similarity: Compositional prefix verbs should occur in the same contexts as their stems
  - poor coverage in typical distributional similarity databases
  - use 10-million-phrase clustering [Lin et al. '10]
- Frequencies of prefix-verb and stem
- Orthographic features for the prefix, stem, whether hyphenated, etc.

Data

- Supplementary data: list of prefixes, stems (acquired semi-automatically)
- Source data: all prefix verbs that occur >1 time in the NYT section of Gigaword
- Human annotation: annotate approx. 1700 prefix verbs in context, $\kappa = 0.82$
  - Evaluation data publicly available (see paper)

Vs. Conventional Morphology

- Compare to CELEX, a comprehensive morphological dictionary
  - Only 39% in CELEX at all

		CELEX Segmentation	
		Split	**Don't Split**
Compositionality	**Yes**	**227**	**10**
Annotation	**No**	**250**	**183**

Vs. Conventional Morphology

- Compare to CELEX, a comprehensive morphological dictionary
  - Only 39% in CELEX at all

		CELEX Segmentation	
		Split	**Don't Split**
Compositionality	**Yes**	**227**	**10**
Annotation	**No**	**250**	**183**

remarry

Vs. Conventional Morphology

- Compare to CELEX, a comprehensive morphological dictionary
  - Only 39% in CELEX at all

		CELEX Segmentation	
		Split	Don't Split
Compositionality	Yes	227	10
Annotation	No	250	183

cooperate

Vs. Conventional Morphology

- Compare to CELEX, a comprehensive morphological dictionary
  - Only 39% in CELEX at all

		CELEX Segmentation	
		Split	Don't Split
Compositionality	Yes	227	10
Annotation	No	250	183

understand

Vs. Conventional Morphology

- Compare to CELEX, a comprehensive morphological dictionary
  - Only 39% in CELEX at all

		CELEX Segmentation	
		Split	**Don't Split**
Compositionality	**Yes**	**227**	**10** await
Annotation	**No**	**250**	**183**

Systems

1. **Base1:** always choose compositional
2. **Base2:** for each prefix, choose majority decision for verbs with that prefix
3. **Morf:** Morfessor system

   [Creutz & Lagos ’07]
4. **SuperComp:** The supervised compositionality detection classifier

Main Results

- Accuracy of classifications on test set (1000 train, 360 test verbs):

Base1	Base2	Morf	SuperComp
65.7%	**87.2%**	**73.8%**	***93.6%***

- Best features: 1) *orthographic*, 2) *hyph-enation*, 3) *N-gram co-occurrence*

More Results

- Performance of systems on prefix verbs that are NOT in CELEX dictionary

  (i.e. rare verbs)

Base1	Base2	Morf	SuperComp
85.3%	**94.8%**	**86.6%**	***95.7%***

Conventional Performance

- Train and test system on CELEX segmentations (506 train, 250 test verbs)

Base1	Base2	Morf	SuperComp
76.0%	**79.6%**	**72.4%**	***86.4%***

- BUT: targeting conventional morphology may be both harder and less useful than targeting semantic compositionality

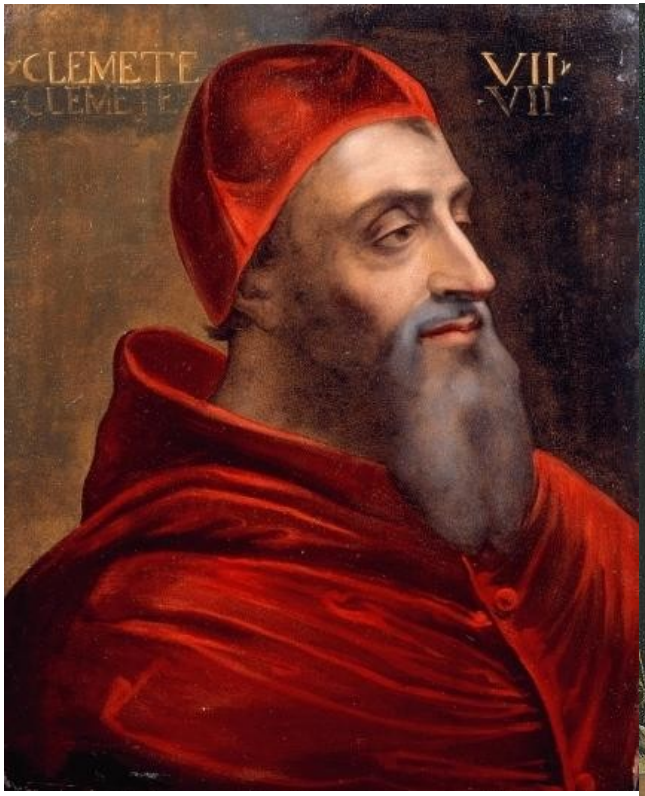
Conclusion

- A new, well-defined and practical definition of compositionality for prefix verbs
- Predict compositionality using a range of orthographic and web-scale statistical features
- Achieve 94% accuracy
- Also performs well on conventional morphological segmentations

Future Work

- New features:
  - lexical fixedness [Fazly et al. '09] of prefix verb:
    - *quest again (***re*****quest**) vs. marry again (***re*****marry**)
  - measure cohesion of stem with context terms
    - "A customer ***re*****quested** a refund" and "**quest**"
- Jointly learn compositionality across inflections
- Applications

Thanks



Porter stemming

Search

About 273,000 results (0.20 seconds)

Advanced search

Everything

More

The web

Pages from Canada

More search tools

Porter Stemming Algorithm

The **Porter stemming** algorithm (or '**Porter stemmer**') is a process for removing the commoner morphological and inflexional endings from words in English. ...

tartarus.org/~martin/**PorterStemmer**/ - Cached - Similar

Stemming - Wikipedia, the free encyclopedia

A later **stemmer** was written by Martin **Porter** and was published in the July 1980 issue of the Journal Program. This **stemmer** was very widely used and became ...

Examples - History - Algorithms - Language Challenges

en.wikipedia.org/wiki/**Stemming** - Cached - Similar

What is **Porter Stemming**?

The **Porter Stemmer** is a conflation Stemmer developed by Martin Porter at the University of Cambridge in 1980. The Stemmer is based on the idea that the ...

www.comp.lancs.ac.uk/computing/.../**stemming**/.../**porter**.htm - Cached - Similar

Porter's Stemming Algorithm Online

Porter's Stemming Algorithm Online. Enter a sequence of words in the box below to stem (Note: "stop" words and punctuation are automatically removed)

maya.cs.depaul.edu/classes/ds575/**porter**.html - Cached

Porter-Stemmer | drupal.org

17 Dec 2005 ... This module implements the **Porter stemming** algorithm to improve English language searching with the Drupal built-in Search module. ...

drupal.org › Download › Modules - Cached - Similar

PHP Class: **Porter Stemming** Algorithm: chuggnutt.com

Free PHP implementation of the **Porter** Stemmimg Algorithm. chuggnutt.com is the personal domain/site of Jon Abernathy, and is devoted to web culture, ...

www.chuggnutt.com/**stemmer**.php - Cached - Similar