Predicting the relevance of distributional semantic similarity with contextual information

Philippe Muller, Cécile Fabre, Clémentine Adam

IRIT & CLLE, University of Toulouse

June 23rd, 2014

supported by Asfalda project (ANR-12-CORD-023)

The big picture

- distributional similarity hypothesis:
  lexical items with same usage contexts are semantically related
- exhibited relations are either classical synonymy/hypernymy or indicative of topical cohesion
- but
  - polysemy and context: relatedness depends on the context of use
  - validation is not obvious

our contribution

- context helps validation of lexical associations
- the relevance of relatedness in a given context can be predicted
- contextual features play a role, not just strength of *a priori* similarity

The distributional hypothesis

Similar contexts (graphic or syntactic):

Pair	contexts	relation
immortal, eternal	soul, love	synonymy
win, lose	game, time, bet	antonymy
apple, fruit	_ salad, harvest _	hyponymy
professor_of, teach_obj	literature, science	?
bottle, die	throw	noise ?

semantic relatedness of words is difficult to assess when presented out of context

Overview

- judging of relatedness : context helps reliability
  - out of context human annotation
  - in context human annotation
- experiment: predicting contextual similarity
  - data
  - setup / methods
  - results

Judging relatedness

- intrinsic relatedness: judging pairs without any context, eg are the items **root,insect** semantically related ?
- contextual relatedness: judging pairs in a text where they appear together: is the relation between the items relevant for textual cohesion ?

  While they are insectivores, hedgehogs are in practice omnivores. They can eat **insects**, **roots**, *melon or squash.*

We must control whether:

- context biases towards relatedness
- annotators can agree on the relevance of item pairs

Judging relatedness

Setup:

- example texts from Wikipedia
- distributional similarity database : French distributional neighbours from `http://redac.univ-tlse2.fr/applications/vdw.html`. built from a 250M word Wikipedia dump, Lin similarity measure from (Lin, 1998).
- linguist annotators are given 100 randomly selected pairs, with additional constraints
  - out-of-context pairs were selected with a minimal similarity score of 0.2 (top 14% of all lexical pairs).
  - in context: no threshold on similarity score, but must appear in the same paragraph somewhere in the corpus
- Then two of the annotators (now “experts”) went on to annotate a larger sample ($\approx$ 2000 pairs).

Example: impala

[...] Le ventre de l'impala de même que ses lèvres et sa queue sont blancs. Il faut aussi mentionner leurs lignes noires uniques à chaque individu au bout des oreilles, sur le dos de la queue et sur le front. Ces lignes noires sont très utiles aux impalas puisque ce sont des signes qui leur permettent de se reconnaitre entre eux. Ils possèdent aussi des glandes sécrétant des odeurs sur les pattes arrières et sur le front. Ces odeurs permettent également aux individus de se reconnaitre entre eux. Il a également des coussinets noirs situés, à l'arrière de ses pattes. Les impalas mâles et femelles ont une morphologie différente. En effet, on peut facilement distinguer un mâle par ses cornes en forme de S qui mesurent de 40 à 90 cm de long.

Les impalas vivent dans les savanes où l' herbe (courte ou moyenne) abonde. Bien qu'ils apprécient la proximité d'une source d'eau, celle-ci n'est généralement pas essentielle aux impalas puisqu'ils peuvent se satisfaire de l'eau contenue dans l' herbe qu'ils consomment. Leur environnement est relativement peu accidenté et n'est composé que d' herbes, de buissons ainsi que de quelques arbres.
[...]

target item: *corne*, horn, in blue
candidates: pending yellow words (*oreille/queue*, ear/tail),
relevant: green words *pattes*, legs)
not relevant: red words: *herbe*, grass

Inter-annotator agreement

N_i : linguists. 2×100 pairs

Annotators	Non-contextual		Contextual	
	Agreement rate	Kappa	Agreement rate	Kappa
N1+N2	77%	0.52	91%	0.66
N1+N3	70%	0.36	92%	0.69
N2+N3	79%	0.50	92%	0.69
Average	75,3%	0.46	91,7%	0.68
Experts	NA	NA	90.8%	0.80

Experts: N1+N2, 2000 pairs.

Predicting relatedness

- Data: the 2000 annotated pairs of lexical items appearing in a common paragraph (either “relevant” or “not relevant”)
- An imbalance classification problem: 11% of pairs only are relevant
- Features:
  - a group for corpus frequencies of target lexical items
  - a group for distributional association related measures
  - a group for contextual information
- Baseline: just use Lin’s score, with a threshold for relevance determined on a sample of the instances.
- Classifiers: Random Forest and Naive Bayes
- Class imbalance addressed with resampling (Smote), and cost-aware learning (MetaCost)
- Evaluation: 10 fold cross validation

Features: corpus

for each pair of lexical items (a, b) ("neighbours"), considering corpus frequencies:

Feature	Description
$freq_{\min}$	$\min(freq_a, freq_b)$
$freq_{\max}$	$\max(freq_a, freq_b)$
$freq_{\times}$	$\log(freq_a \times freq_b)$
mi	$mi = \log \frac{P(a,b)}{P(a) \cdot P(b)}$

Features: distributional similarity

given similarity scores, we can use rankings of items similar to another one, and productivity of items (the number of times they appear as similar to another item)

Feature	Description
lin	Lin's score
$rank_{\min}$	$\min(rank_{a-b}, rank_{b-a})$
$rank_{\max}$	$\max(rank_{a-b}, rank_{b-a})$
$rank_{\times}$	$\log(rank_{a-b} \times rank_{b-a})$
$prod_{\min}$	$\min(prod_a, prod_b)$
$prod_{\max}$	$\max(prod_a, prod_b)$
$prod_{\times}$	$\log(prod_a \times prod_b)$
cats	neighbour pos pair (eg NN, AN,...)
predarg	predicate or argument

Features: context

given a set of occurrences of item a and b in the same text,
use frequencies in this context, distances between occurrences,
related items (productivity within the text, connected components)

Feature	Description
$freqtxt_{\min}$	$\min(freqtxt_a, freqtxt_b)$
$freqtxt_{\max}$	$\max(freqtxt_a, freqtxt_b)$
$freqtxt_{\times}$	$\log(freqtxt_a \times freqstxt_b)$
$tf{\cdot}ipf$	$tf{\cdot}ipf(\text{neighbour}_a) \times tf{\cdot}ipf(\text{neighbour}_b)$
$copr_{ph}, copr_{para}$	copresence in a sentence, paragraph
sd, gd, ad	smallest, highest, average distance between neighbour_a and neighbour_b
$prodtxt_{\min}, max$	$\min(prod_a, prod_b), max(...)$
$prodtxt_{\times}$	$\log(prod_a \times prod_b)$
cc	belong to the same lexical connected component

Results

Method	Precision	Recall	F-score	CI
Baseline (Lin threshold)	24.0	24.0	24.0	
RF	**68.1**	24.2	35.7	$\pm$ 3.4
NB	34.8	51.3	41.5	$\pm$ 2.6
RF+resampling	56.6	32.0	40.9	$\pm$ 3.3
NB+resampling	32.8	54.0	40.7	$\pm$ 2.5
RF+cost aware learning	40.4	54.3	**46.3**	$\pm$ 2.7
NB+cost aware learning	27.3	**61.5**	37.8	$\pm$ 2.2

Feature Impact

Features	Prec.	Recall	F-score
all	40.4	54.3	46.3
all − corpus feat.	37.4	52.8	43.8
all − similarity feat.	36.1	49.5	41.8
all − contextual feat.	36.5	54.8	43.8

Take-away

- distributional similarity hypothesis:
  lexical items with same usage contexts are semantically related
- exhibited relations are either classical synonymy/hypernymy or indicative of topical cohesion
- but
  - polysemy and context: relatedness depends on the context of use
  - validation is not obvious

our contribution

- context helps validation of lexical associations
- the relevance of relatedness in a given context can be predicted
- contextual features play a role, not just strength of *a priori* similarity