Disambiguating Distributional Neighbors using a Lexical Substitution Dataset

François Morlane-Hondère, Cécile Fabre,
Nabil Hathout and Ludovic Tanguy

CLLE-ERSS, University of Toulouse & CNRS

NLPCS 2014, Venice
October 27th 2014

Outline

Outline

Distributional semantics

- The distributional hypothesis (Harris 1954):

  DISTRIBUTIONAL SIMILARITY $\Leftrightarrow$ SEMANTIC SIMILARITY

- Distributional similarity computation:
  1. extraction of words' contexts from a corpus
  2. weighting of the contexts
  3. calculation of the distributional similarity between words

- Many models. . . (Distributional Memory (Baroni & Lenci 2010), LSA (Deerwester et al. 1990), HAL (Lund et al. 1995) etc.)

. . . for many applications (information retrieval, thesaurus construction, question answering, linguistic research, etc.)

Distributional semantics

- The distributional hypothesis (Harris 1954):

  DISTRIBUTIONAL SIMILARITY $\Leftrightarrow$ SEMANTIC SIMILARITY

- Distributional similarity computation:
  1. extraction of words' contexts from a corpus
  2. weighting of the contexts
  3. calculation of the distributional similarity between words

- Many models. . . (Distributional Memory (Baroni & Lenci 2010), LSA (Deerwester et al. 1990), HAL (Lund et al. 1995) etc.)

. . . for many applications (information retrieval, thesaurus construction, question answering, linguistic research, etc.)

Distributional semantics

- The distributional hypothesis (Harris 1954):

  DISTRIBUTIONAL SIMILARITY $\Leftrightarrow$ SEMANTIC SIMILARITY

- Distributional similarity computation:
  1. extraction of words' contexts from a corpus
  2. weighting of the contexts
  3. calculation of the distributional similarity between words

- Many models... (Distributional Memory (Baroni & Lenci 2010), LSA (Deerwester et al. 1990), HAL (Lund et al. 1995) etc.)

...for many applications (information retrieval, thesaurus construction, question answering, linguistic research, etc.)

A first example

- 10 closest neighbors of *conference* in Distributional Memory

*meeting*_N	0.838
*seminar*_N	0.823
*symposium*_N	0.806
*rally*_N	0.725
*congress*_N	0.711
*workshop*_N	0.699
*colloquium*_N	0.684
*gathering*_N	0.667
*event*_N	0.650
*fair*_N	0.648

A second example

- 10 closest neighbors of *mouse* in Distributional Memory

*rat*_N	0.542
*animal*_N	0.466
*rabbit*_N	0.449
*cursor*_N	0.440
*monkey*_N	0.424
*cat*_N	0.421
*pig*_N	0.413
*joystick*_N	0.384
*dog*_N	0.375
*human*_N	0.373

A second example

- 10 closest neighbors of *mouse* in Distributional Memory

→ mix of **ANIMAL** and **DEVICE** meaning-related neighbors

*rat*_N	0.542
*animal*_N	0.466
*rabbit*_N	0.449
*cursor*_N	0.440
*monkey*_N	0.424
*cat*_N	0.421
*pig*_N	0.413
*joystick*_N	0.384
*dog*_N	0.375
*human*_N	0.373

The ambiguity problem

- Distributional thesauri: reflection of words' distributions
- But words can be ambiguous

The ambiguity problem

- Distributional thesauri: reflection of words' distributions

- But words can be ambiguous

for a decent breeder. get in touch with the local rat and	mouse	club before getting pet mice. look in the local papers as well
Sala de la Resurreccion : rotate the image using the	mouse	. Found in the first days of the 1990 expedition , this
(30 OCT 74) Of Mice And Women There 's a	mouse	loose in the flat . Robin sees a way of taking advantage
Jezierski (KocieSymbole , containing cat and	mouse	dingbats ? larger illustrations of this font on the creator 's site
be calcified and shows up on an XRay as a " joint	mouse	" . Later osteoarthritis (also called degenerative joint disease)
consist of any number of people . 2. No more than 2	mouse	handlers are allowed to operate the mouse during the competition
change by clicking and dragging over them with your	mouse	, just like you would select a range of cells in Microsoft
pace . Each step can be completed in no more than 4	mouse	click and as little as 1 due to automatic item selection and
the line , over the white square box Hold down your left	mouse	button when the mouse pointer changes to the shape of the
will show only the jobs you have sent to print Use the	mouse	to select the document you wish to print - NB : your
' Stuart Little ' the story of Stuart Little , an adorable	mouse	who is adopted by Mr and Mrs Little as a brother for

ukWaC corpus (Baroni et al. 2009)

The ambiguity problem

- Data from Distributional Memory

	*trap*_V OBJ	*eat*_V SUBJ	*driver*_N MOD	*plug*_V OBJ	...
*rodent*_N	13.85	24.33	0	0	...
*rat*_N	14.08	222.48	0.33	0	...
*animal*_N	170.06	511.04	0	0	...
*device*_N	4.27	0.06	78.77	254.62	...
*keyboard*_N	0	0	29.33	97.13	...
*joystick*_N	0	0	38.92	10.06	...
⋮	⋮	⋮	⋮	⋮	⋱

*mouse*_N	24.89	148.23	52.17	53.94	...

$\Rightarrow$ **heterogeneous context vectors**

The ambiguity problem

- Data from Distributional Memory

	*trap*_V OBJ	*eat*_V SUBJ	*driver*_N MOD	*plug*_V OBJ	...
*rodent*_N	13.85	24.33	0	0	...
*rat*_N	14.08	222.48	0.33	0	...
*animal*_N	170.06	511.04	0	0	...
*device*_N	4.27	0.06	78.77	254.62	...
*keyboard*_N	0	0	29.33	97.13	...
*joystick*_N	0	0	38.92	10.06	...
⋮	⋮	⋮	⋮	⋮	⋱

*mouse*_N	24.89	148.23	52.17	53.94	...

$\Rightarrow$ **heterogeneous context vectors**

Outline

The word sense disambiguation (WSD) framework

Differences with a traditional WSD system:

- no contextual information
- no sense-tagged corpus
- no WordNet-like sense repository

Our framework:

- a non-annotated corpus
- a lexical substitution dataset
- a distributional thesaurus

The word sense disambiguation (WSD) framework

Differences with a traditional WSD system:

- no contextual information
- no sense-tagged corpus
- no WordNet-like sense repository

Our framework:

- a non-annotated corpus
- a lexical substitution dataset
- a distributional thesaurus

The SemDis 2014 lexical substitution dataset

- Gold standard of the SemDis 2014 French lexical substitution task (Fabre et al. 2014)

- Goal of the task: to develop systems that give the best substitutes for a set of contextualized target words

- Example from the SemEval-2007 English Lexical Substitution Task (McCarthy and Navigli, 2009):

  The ideal preparation would be a light meal about 2-2 1/2 hours pre-match, followed by a warm-up hit and perhaps a top-up with extra fluid before the ***match****.*

The SemDis 2014 lexical substitution dataset

- Need for an evaluation gold standard
- Development of the dataset:
  1. selection of 30 French polysemous target words (N, V and A)
  2. selection of context sentences in frWaC:
     - 10 sentences for each target word
     - each sentence is associated to one of the target word's senses
  3. human annotation: annotators – 7 per sentence – were asked to provide up to 3 substitutes for each target word
- Freely-available (`http://www.irit.fr/semdis2014/fr/task1.html`)

The SemDis 2014 lexical substitution dataset

- Substitutes provided for the word *affection* 'affection' (ILLNESS vs LOVE):
  - *affection* as ILLNESS: *maladie* 'illness' (22), *trouble* 'disorder' (9), *pathologie* 'pathology' (7), *mal* 'ache' (5), *dysfonctionnement* 'dysfunction' (3), *atteinte* 'harm' (2), *problème* 'problem' (2), *anomalie* 'anomaly' (2), *condition* 'condition' (1)...
  - *affection* as LOVE: *amour* 'love' (26), *tendresse* 'tenderness' (16), *attachement* 'attachment' (6), *amitié* 'friendship' (4), *sentiment* 'feeling' (4), *attention* 'attentions' (2), *lien* 'bond' (2), *intimité* 'intimacy' (1), *sollicitation* 'appeal' (1), *proximité* 'closeness' (1)...

The SemDis 2014 lexical substitution dataset

- Properties of the dataset:
  - the coverage is very limited
  - the development of such a dataset is time-consuming

  BUT
  - each substitute is related to a sense
  - the different senses are easily distinguished from one another
  - the semantic relatedness is valued (number of annotators)
  - all the substitutes are commonly used words and are thus frequent in contemporary corpora
- Value of this resource for disambiguating distributional neighbors?

The SemDis 2014 lexical substitution dataset

- Properties of the dataset:
  - the coverage is very limited
  - the development of such a dataset is time-consuming

  BUT

  - each substitute is related to a sense
  - the different senses are easily distinguished from one another
  - the semantic relatedness is valued (number of annotators)
  - all the substitutes are commonly used words and are thus frequent in contemporary corpora

- Value of this resource for disambiguating distributional neighbors?

The SemDis 2014 lexical substitution dataset

- Properties of the dataset:
  - the coverage is very limited
  - the development of such a dataset is time-consuming

  BUT

  - each substitute is related to a sense
  - the different senses are easily distinguished from one another
  - the semantic relatedness is valued (number of annotators)
  - all the substitutes are commonly used words and are thus frequent in contemporary corpora

- Value of this resource for disambiguating distributional neighbors?

The SemDis 2014 lexical substitution dataset

- Selection of a subset of 11 out of the 30 SemDis two-meaning target words
  - 4 nouns: *affection* (LOVE vs ILLNESS), *débit* (DEBIT vs OUTPUT), *don* (TALENT vs PRESENT), *montée* (RISE vs ASCENT)
  - 4 verbs: *entraîner* (TRAIN vs LEAD TO), *fonder* (CREATE vs BASE ON), *interpréter* (UNDERSTAND vs PERFORM), *maintenir* (HOLD vs ASSERT)
  - 3 adjectives: *aisé* (EASY vs RICH), *mince* (THIN vs POOR), *riche* (RICH vs WEALTHY)

The distributional thesaurus

- Properties:
  - generated from a 262-million words corpus of French Wikipedia articles
  - contains the nouns, verbs and adjectives which occur in at least 5 different contexts in the corpus
  - syntactic dependencies were used as contexts using the Talismane parser (Urieli, 2013)
  - weighting of the contexts: pointwise mutual information
  - similarity measure: cosine measure

- The distributional thesaurus is used to provide a list of 10 distributional neighbors for each of the 11 target words

The distributional thesaurus

- Properties:
  - generated from a 262-million words corpus of French Wikipedia articles
  - contains the nouns, verbs and adjectives which occur in at least 5 different contexts in the corpus
  - syntactic dependencies were used as contexts using the Talismane parser (Urieli, 2013)
  - weighting of the contexts: pointwise mutual information
  - similarity measure: cosine measure

- The distributional thesaurus is used to provide a list of 10 distributional neighbors for each of the 11 target words

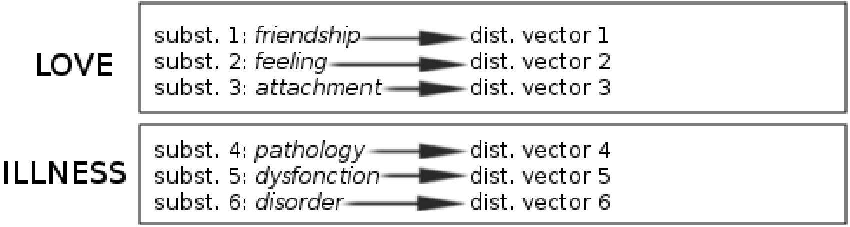
How to find the meaning of a target word to which its neighbors are related?

1. Generation of the *sense vectors* of *affection* (Wiki. corpus)

LOVE	subst. 1: *friendship* subst. 2: *feeling* subst. 3: *attachment*
ILLNESS	subst. 4: *pathology* subst. 5: *dysfonction* subst. 6: *disorder*

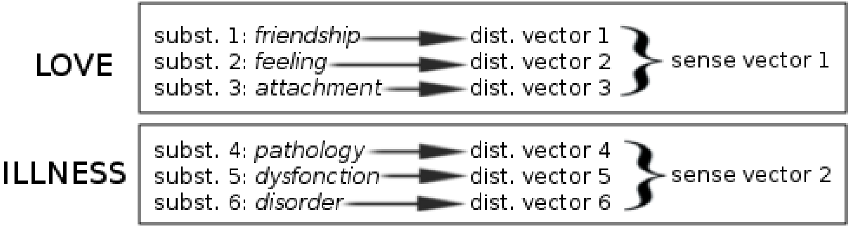
How to find the meaning of a target word to which its neighbors are related?

1. Generation of the *sense vectors* of *affection* (Wiki. corpus)



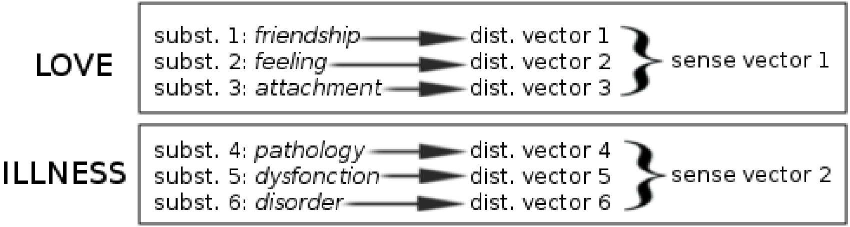
How to find the meaning of a target word to which its neighbors are related?

1. Generation of the *sense vectors* of *affection* (Wiki. corpus)



How to find the meaning of a target word to which its neighbors are related?

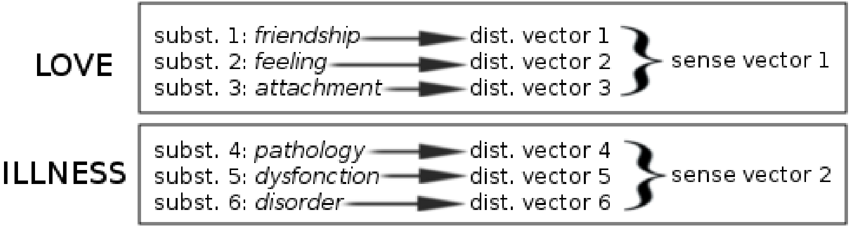
1. Generation of the *sense vectors* of *affection* (Wiki. corpus)



2. Generation of the neighbors' distributional vectors

How to find the meaning of a target word to which its neighbors are related?

1. Generation of the *sense vectors* of *affection* (Wiki. corpus)



2. Generation of the neighbors' distributional vectors
3. Measure of the similarity between neighbor's vectors and sense vectors

Method

- Results of the similarity measure for the neighbors of the target word *affection*

	senses	
	LOVE	ILLNESS
complication 'complication'	0.110	**0.590**
lésion 'lesion'	0.086	**0.744**
sympathie 'sympathy'	**0.680**	0.062
admiration 'admiration'	**0.776**	0.079
infection 'infection'	0.126	**0.691**
tumeur 'tumor'	0.052	**0.520**
symptôme 'symptom'	0.228	**0.542**
estime 'esteem'	**0.300**	0.024
manifestation 'manifestation'	0.134	**0.603**
épilepsie 'epilepsy'	0.083	**0.324**

Evaluation

- Manual annotation of the 10 nearest neighbors for each of the 11 SemDis target words (25 % pariwise agreement)
- In the annotation guidelines, a neighbor can be related to:
  - the 1^{st} or the 2^{nd} sense of the target word
    → 70 % of the 110 neighbours
  - both senses
    → 10 %
  - none
    → 20 %

Evaluation

- Need to compare the similarity scores and the annotations
- Development of a rule-based decision system based on 2 variables:
  - the ratio between the two similarity values (threshold σ_r)
  - the lowest of the two similarity values (threshold σ_{low})
- Illustration with the target word *débit* 'debit'

	FLUX 'FLOW'	RETRAIT 'WITHDRAWAL'	
taux 'rate'	0,747	0,511	sense 1
revenu 'income'	0,529	0,521	both
précipitation 'precipitations'	0,395	0,419	none

Evaluation

- Need to compare the similarity scores and the annotations
- Development of a rule-based decision system based on 2 variables:
  - the ratio between the two similarity values (threshold σ_r)
  - the lowest of the two similarity values (threshold σ_{low})
- Illustration with the target word *débit* 'debit'

	FLUX 'FLOW'	RETRAIT 'WITHDRAWAL'	
taux 'rate'	0,747	0,511	sense 1
revenu 'income'	0,529	0,521	both
précipitation 'precipitations'	0,395	0,419	none

Results

- Accuracy of 0.64 over the 110 neighbors

Gold \ System	None	Both	Sense 1	Sense 2	Total
None	5	0	2	15	22
Both	0	1	8	2	11
Sense 1	1	0	36	0	37
Sense 2	4	3	4	29	40
Total	10	4	50	46	110

- Confusions with the categories *None* and *Both*

- Only 4 cases out of 110 where the system chose one sense and the annotators chose the other

Results

- Unequal performance among the 11 target words
  - contexts of the sense vectors of *affection* (accuracy 10/10)

ILLNESS	LOVE
*neurologique*_A 'neurologic'_A	*éprouver*_V 'feel'_V
*chronique*_A 'chronic'_A	*profond*_A 'deep'_A
*mental*_A 'mental'_A	*sentiment*_N 'feeling'_N
*soigner*_V 'cure'_V	*paternel*_A 'paternal'_A
*grave*_A 'severe'_A	*vouer*_V 'vow'_V

  - contexts of the sense vectors of *fonder* 'to found' (accuracy 4/10)

CREATE	BASE ON
*fait*_N 'fact'_N	*occasion*_N 'occasion'_N
*principe*_N 'principle'_N	*critère*_N 'criterion'_N
*idée*_N 'idea'_N	*modèle*_N 'model'_N
*décision*_N 'decision'_N	*emplacement*_N 'place'_N
*hypothèse*_N 'hypothesis'_N	*remplacement*_N 'replacement'_N

→ **difference of distance between the sense vectors**

Results

- Unequal performance among the 11 target words
  - contexts of the sense vectors of *affection* (accuracy 10/10)

ILLNESS	LOVE
*neurologique*_A 'neurologic'_A	*éprouver*_V 'feel'_V
*chronique*_A 'chronic'_A	*profond*_A 'deep'_A
*mental*_A 'mental'_A	*sentiment*_N 'feeling'_N
*soigner*_V 'cure'_V	*paternel*_A 'paternal'_A
*grave*_A 'severe'_A	*vouer*_V 'vow'_V

  - contexts of the sense vectors of *fonder* 'to found' (accuracy 4/10)

CREATE	BASE ON
*fait*_N 'fact'_N	*occasion*_N 'occasion'_N
*principe*_N 'principle'_N	*critère*_N 'criterion'_N
*idée*_N 'idea'_N	*modèle*_N 'model'_N
*décision*_N 'decision'_N	*emplacement*_N 'place'_N
*hypothèse*_N 'hypothesis'_N	*remplacement*_N 'replacement'_N

→ **difference of distance between the sense vectors**

Outline

Conclusion

- Problem of ambiguity of distributional neighbors
- Is a lexical substitution dataset of any help?
  - ✗ not a large-scale solution
    - very limited coverage
    - lot of manual intervention
  - ✓ an interesting alternative to WordNet-like thesauri
  - ✓ a harder task for some target words
    → linguistic interest: a closer look to the distributional mechanisms

Conclusion

- Problem of ambiguity of distributional neighbors

- Is a lexical substitution dataset of any help?

  ✗ not a large-scale solution
  - very limited coverage
  - lot of manual intervention

  ✓ an interesting alternative to WordNet-like thesauri

  ✓ a harder task for some target words
  → linguistic interest: a closer look to the distributional mechanisms

Conclusion

- Problem of ambiguity of distributional neighbors
- Is a lexical substitution dataset of any help?
  - ✗ not a large-scale solution
    - very limited coverage
    - lot of manual intervention
  - ✓ an interesting alternative to WordNet-like thesauri
  - ✓ a harder task for some target words
    → linguistic interest: a closer look to the distributional mechanisms

Conclusion

- Problem of ambiguity of distributional neighbors

- Is a lexical substitution dataset of any help?
  ✗ not a large-scale solution
    - very limited coverage
    - lot of manual intervention

  ✓ an interesting alternative to WordNet-like thesauri

  ✓ a harder task for some target words
  → linguistic interest: a closer look to the distributional mechanisms

References

- Baroni, M., Bernardini, S., Ferraresi, A., Zanchetta, E. (2009). The WaCky Wide Web: A collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, 43(3):209–226.
- Baroni, M. and Lenci, A. (2010). Distributional Memory: A general framework for corpus-based semantics. *Computational Linguistics*, 36(4): 673–721.
- Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., Harshman, R. (1990). Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science*, 41(6): 391–407.
- Harris, Z. (1954). Distributional structure. *Word*, 10(23): 146–162.
- Lund, K., Burgess, C., Atchley, R. A. (1995). Semantic and associative priming in a high-dimensional semantic space. *Cognitive Science Proceedings*, 660–665.
- McCarthy, D., Navigli, R. (2009). The English lexical substitution task. *Language Resources and Evaluation*, 43(2): 139–159.
- Urieli, A. (2013). *Robust French syntax analysis: reconciling statistical methods and linguistic knowledge in the Talismane toolkit*. Phd thesis, Univ. of Toulouse