From Non Word to New Word: Automatically Identifying Neologisms in French Newspapers

Ingrid Falk Delphine Bernhard Christophe Gérard

Linguistique, langues, parole – Université de Strasbourg

Language Resources and Evaluation Conference, 2014

Summary

What this talk is about I

Neologisms. . .

How to automatically detect and document them in French online newspaper articles?

Background: the Logoscope system

- Retrieves French newspaper articles on a daily basis,
- Identifies unknown words using lists of known words for the French language (exclusion lists)
- Unknown words are presented to a linguist expert

Background: the Logoscope system

- Retrieves French newspaper articles on a daily basis,
- Identifies unknown words using lists of known words for the French language (exclusion lists)
- Unknown words are presented to a linguist expert

Background: the Logoscope system

- Retrieves French newspaper articles on a daily basis,
- Identifies unknown words using lists of known words for the French language (exclusion lists)
- Unknown words are presented to a linguist expert

Problem

Examples of unknown words in online articles, sorted by frequency:

lmd (18)	twitter/widgets (7)	india-mahdavi (3)
pic(this (18)	garde-à (6)	kilomètresc (2)
lazy-retina (9)	ex-PPR (4)	geniculatus (2)
onload (9)	pro-Morsi (4)	margin-bottom (2)
onerror (9)	tuparkan (4)	politique»(2)
amp;euro (7)	candiudature (3)	. . .

Table: The most frequent unknown words collected on 2013-07-12. In parentheses: frequency.

What this talk is about II

- Select among unknown words the most probable neologism candidates
- Using a classification method
- Which features are most helpful?

What this talk is about II

- Select among unknown words the most probable neologism candidates
- Using a classification method
- Which features are most helpful?

What this talk is about II

- Select among unknown words the most probable neologism candidates
- Using a classification method
- Which features are most helpful?

Summary

Method

- Collect corpus from French newspaper RSS feeds
- Filter based on exclusion list
- Classify resulting unknown words:

  Positive: is a valid neologism
  Negative: is not a valid neologism

Collected Corpus

Total number of articles: 2,723

Newspapers: Le Monde (659), Libération (504), l'Équipe (594), Les Echos (956)

Dates: 7 weekdays in July 2013

Total number of forms (tokens): 51,000

Filtered using exclusion list:

Unknown forms (types): 692

Manually validated:

True neologisms (types): 81

Exclusion list

Morphalou: Morpho-syntactic lexicon for French [Romary et al., 2004]

Wortschatz: Corpus based word list [Biemann et al., 2004]

Named entities: CasEN Named Entity Recognition system [Maurel et al., 2011]

The Classification

Supervised, SVM: LibSVM [Chang and Lin, 2011] and Weka [Hall et al., 2009] implementations

Training/testing data: 692 unknown words, manually validated, 81 neologisms

Features: described on following slides

Evaluation: 10-fold cross-validation

Features

- Form related
- Morpho-lexical
- Thematic

Extracted from the corpus

Form related features

- form of string
- language independent
- Examples:
  - length of string,
  - whether string contains dashes,
  - frequency

Morpho-lexical features

- language dependent

prefixes/suffixes *néopaganisme*, *hollandisme*

language cues *pinterest* $\xrightarrow{\texttt{Lingua::Identify}}$ is English with 0.015 probability

spelling *stupédiants* $\xrightarrow{\texttt{aspell}}$ Levenshtein distance to form in dictionary $= 1$

composite? *présidentdirecteur* $\xrightarrow{\texttt{Aho-Corasick}}$ longest substring in known forms has length 9

Morpho-lexical features

- language dependent

prefixes/suffixes *néopaganisme*, *hollandisme*

language cues *pinterest* $\xrightarrow{\texttt{Lingua::Identify}}$ is English with 0.015 probability

spelling *stupédiants* $\xrightarrow{\texttt{aspell}}$ Levenshtein distance to form in dictionary $= 1$

composite? *présidentdirecteur* $\xrightarrow{\texttt{Aho-Corasick}}$ longest substring in known forms has length 9

Morpho-lexical features

- language dependent

prefixes/suffixes *néo*paganisme, holland*isme*

language cues *pinterest* $\xrightarrow{\texttt{Lingua::Identify}}$ is English with 0.015 probability

spelling *stupédiants* $\xrightarrow{\texttt{aspell}}$ Levenshtein distance to form in dictionary $= 1$

composite? *présidentdirecteur* $\xrightarrow{\texttt{Aho-Corasick}}$ longest substring in known forms has length 9

Morpho-lexical features

- language dependent

prefixes/suffixes *néopaganisme*, *hollandisme*

language cues *pinterest* $\xrightarrow{\texttt{Lingua::Identify}}$ is English with 0.015 probability

spelling *stupédiants* $\xrightarrow{\texttt{aspell}}$ Levenshtein distance to form in dictionary $= 1$

composite? *présidentdirecteur* $\xrightarrow{\texttt{Aho-Corasick}}$ longest substring in known forms has length 9

Morpho-lexical features

- language dependent

prefixes/suffixes *néopaganisme*, *hollandisme*

language cues *pinterest* $\xrightarrow{\texttt{Lingua::Identify}}$ is English with 0.015 probability

spelling *stupédiants* $\xrightarrow{\texttt{aspell}}$ Levenshtein distance to form in dictionary $= 1$

composite? *présidentdirecteur* $\xrightarrow{\texttt{Aho-Corasick}}$ longest substring in known forms has length 9

Thematic features

Intuition

New words appear in specific thematic contexts

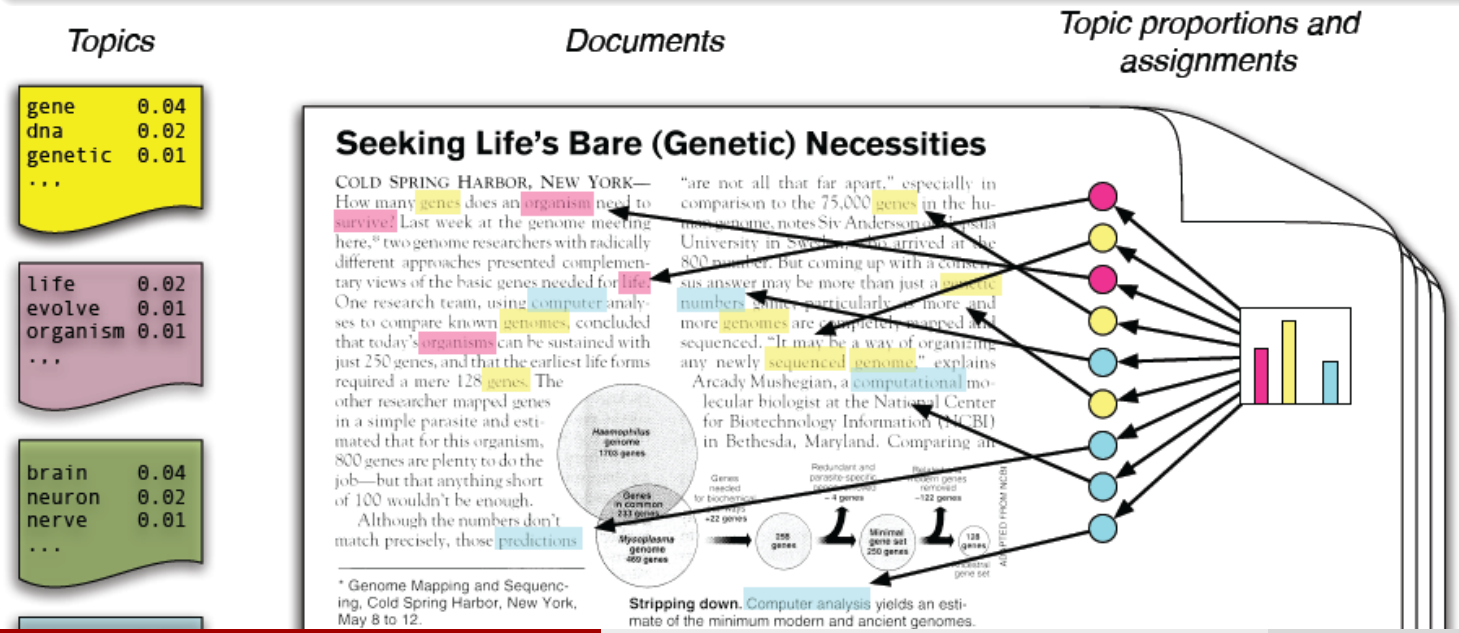
Capture thematic context of new words

- *Newspaper*, *eg.* Le Monde, Libération, etc.
- based on *Topic Modeling*

Topic Modeling

Intuition

- Documents ≈ Mixture of topics
- Topics ≈ Probability distributions over words
- Topic modeling ≈ Use probabilistic graphical model to infer topics from collection of documents



Thematic Features Using Topic Modeling

Learn a topic model of 10 general journalistic topics

- from 4,755 French online articles (*Le Monde*, *Libération*, etc.)
- different from those used in classification experiments

Apply topic model on context of unknown words

- probability distribution over Topic 1 . . . Topic 10

Context 1	concat. of sentences containing unknown word
features:	topic proportions, *eg.*: T1=0.05, . . . , T5=0.54, etc.
Context 2	articles containing unknown word
features:	maximal proportion of Topic *n* found in articles

Thematic Features Using Topic Modeling

Learn a topic model of 10 general journalistic topics

- from 4,755 French online articles (*Le Monde*, *Libération*, etc.)
- different from those used in classification experiments

Apply topic model on context of unknown words

- probability distribution over Topic 1 . . . Topic 10

Context 1	concat. of sentences containing unknown word
features:	topic proportions, *eg.*: T1=0.05, . . . , T5=0.54, etc.
Context 2	articles containing unknown word
features:	maximal proportion of Topic *n* found in articles

Summary

Quantitative Evaluation

7 classifications

7 feature sets: all combinations of formal, morpho-lexical and thematic features.

10-fold cross-validation

- precision, recall and F-measure
  - for positive class
  - averaged over positive and negative class.
- number of validated neologisms (true positives).

Quantitative Results

form, lex, theme				
class	**Prec**	**Rec**	**F**	**corr.**
pos	0.181	0.827	0.297	
both	0.868	0.548	0.625	67

form, lex			
Prec	**Rec**	**F**	**corr.**
0.192	0.778	0.308	
0.864	0.597	0.669	63

form, theme				
class	**Prec**	**Rec**	**F**	**corr.**
pos	0.160	0.531	**0.346**	
both	0.826	0.625	0.693	43

form			
Prec	**Rec**	**F**	**corr.**
0.190	0.481	0.273	
0.832	0.704	**0.752**	39

lex				
class	**Prec**	**Rec**	**F**	**corr.**
pos	0.132	0.827	0.227	
both	0.836	0.350	0.415	67

theme			
Prec	**Rec**	**F**	**corr.**
0.129	**0.889**	0.225	
0.844	0.295	0.338	**72**

lex, theme				
pos			0.136	0.877	0.236	
both			0.851	0.345	0.404	71

Quantitative Results – Interpretation

- best F-measure for global classification task using form features
  - but least identified validated neologisms.
- most identified valid neologisms using theme feature
  - but low global F-measure.
- best balance between global F-measure and detected neologisms using form, lex, theme features

Qualitative Results: reordered sample

ultra-présent (−)	crypto-fascisme (−)	anti-défilé (−)
Etat-département (−)	semi-itinérants (−)	pro-MDC (−)
anti-alcoolisme (−)	mini-Internationale (−)	anti-monégasque (−)
pagano-satanisme (+)	neo-retraité (+)	entraîneur-athlète (−)
watts-étalons (−)	écarts-types (−)	néonicotinoides (−)
auto-diag−stiqués	agroécologiste (+)	...

Table: Unknown words ranked by SVM probability. Classification obtained with form, lex, theme features. In parentheses: if validated or not.

Qualitative Evaluation: Feature Sets

Features (#neos)	**top 5 identified valid neologisms**
form (37)	supermédiateur, doublevédoublevédoublevé, auto-diagnostiqués, néo-célibataires, sur-monétisation ⇝ long, containing '-'
lex (48)	agroécologiste, multiactivité, auto-obscurcissant, neo-retraité, macrostabilité ⇝ compositions, with or w/o prefix
theme (48)	e-détournements, partenadversaires, hollan-disme, retricoté, agnélise ⇝ no salient formal or morpho-lexical proper-ties

Summary

Conclusion

Using an SVM classifier and feature sets based on word form, morpho-lexical characteristics and thematic context

- unknown forms could be reordered in a more meaningful way
- morpho-lexical and thematic features had a considerable contribution in the reordering.

Future Work

Further explore the impact of features.

- Other features:
  - morpho-syntactic properties
  - position of unknown word in text
  - the journalistic genre
  - better topic model
- More precise feature exploration methods: [Lamirel et al., 2013]

Thank you!

Questions?

Summary


Biemann, Christian, Bordag, Stefan, Heyer, Gerhard, Quasthoff, Uwe, and Wolff, Christian.
(2004).
Language-Independent Methods for Compiling Monolingual Lexical Data.
In Goos, Gerhard, Hartmanis, Juris, Leeuwen, Jan, and Gelbukh, Alexander, editors, *Computational Linguistics and Intelligent Text Processing*, volume 2945, pages 217–228. Springer Berlin Heidelberg.


Chang, Chih-Chung and Lin, Chih-Jen.
(2011).
LIBSVM: A library for support vector machines.
ACM Transactions on Intelligent Systems and Technology, 2:27:1–27:27.

Software available at
`http://www.csie.ntu.edu.tw/~cjlin/libsvm`.


Lamirel, Jean-Charles, Cuxac, Pascal, Hajlaoui, Kafil, and Chivukula, Aneesh Sreevallabh.
(2013).
A new feature selection and feature contrasting approach based on quality metric: application to efficient classification of complex textual data.
In *International Workshop on Quality Issues, Measures of Interestingness and Evaluation of Data Mining Models (QIMIE)*, Australie, April.


Hall, Mark, Frank, Eibe, Holmes, Geoffrey, Pfahringer, Bernhard, Reutemann, Peter, and Witten, Ian H.
(2009).
The WEKA data mining software: an update.
SIGKDD Explor. Newsl., 11(1):10–18, November.


Maurel, Denis, Friburger, Nathalie, Antoine, Jean-Yves, Eshkol, Iris, and Nouvel, Damien.
(2011).
Cascades de transducteurs autour de la reconnaissance des entités nommées.
Traitement Automatique des Langues, 52(1):69–96.


Romary, Laurent, Salmon-Alt, Susanne, and Francopoulo, Gil. (2004).
Standards going concrete: from LMF to Morphalou.
In *Workshop Enhancing and Using Electronic Dictionaries*, Geneva, Switzerland.