ISTEX Enrichissements
Extraction terminologique avec TermSuite
http://termsuite.github.io/

Damien Cram, Béatrice Daille

Laboratoire des Sciences du Numérique de Nantes (LS2N - UMR 6004)
damien.cram@univ-nantes.fr, beatrice.daille@univ-nantes.fr

Mardi 6 juin 2017

TermSuite

Un lien

`http://termsuite.github.io/`

Deux fonctionnalités principales

1. Extraction terminologique
2. Alignement multilingue

Trois interfaces

1. API Java
2. Ligne de commande
3. Interface graphique

Six langues

Français, Anglais, Allemand, Russe, Espagnol, Italien

Extraction terminologique
Définitions

Termes

- termes simples (TS) et complexes (TC)
- composés néoclassiques et natifs
- composés syntagmatiques

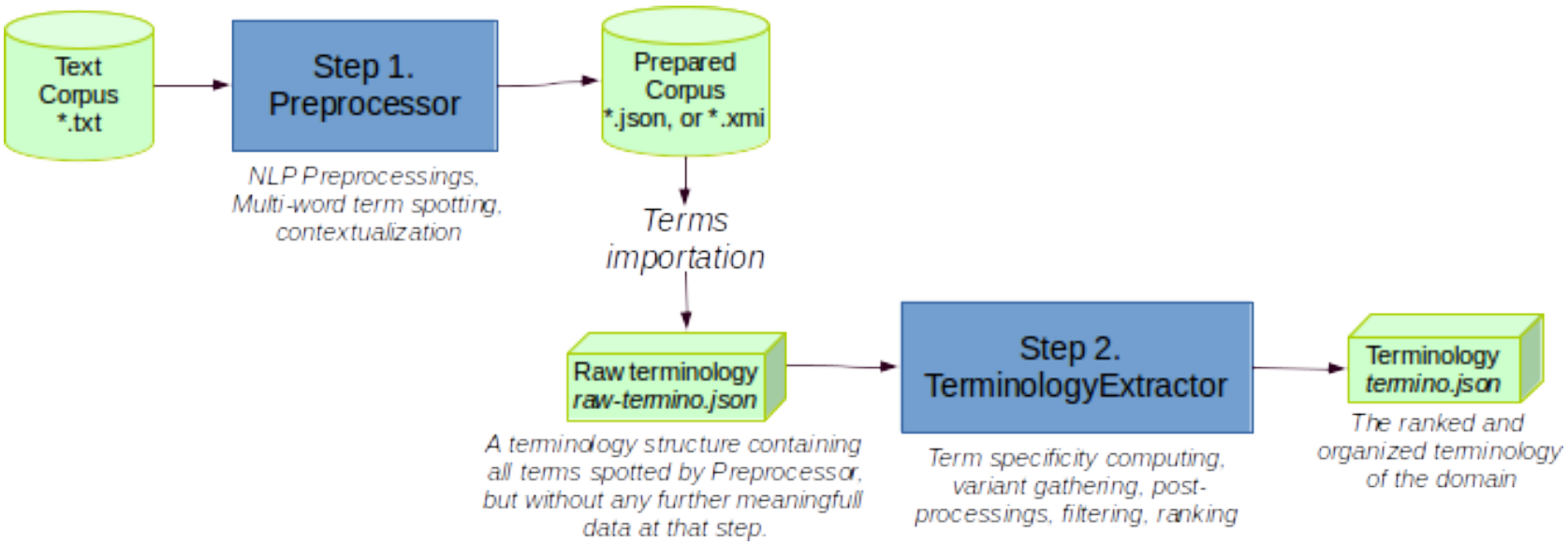
Variantes de termes

2015 variantes syntagmatiques, variantes morphologiques

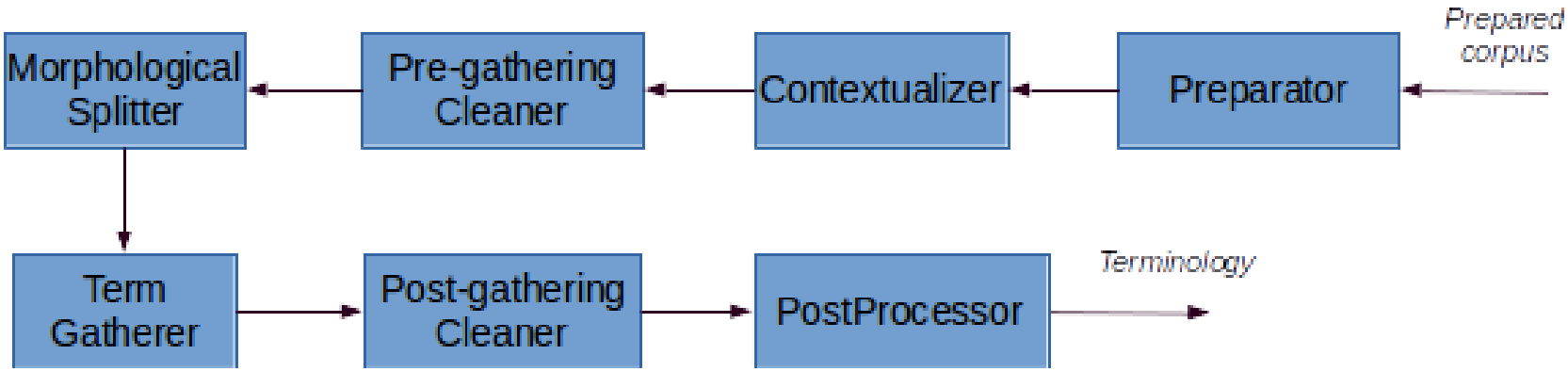
2016 variantes dérivatives, variantes préfixatives

2017 variantes sémantiques

Processus d'extraction terminologique de TermSuite en deux étapes

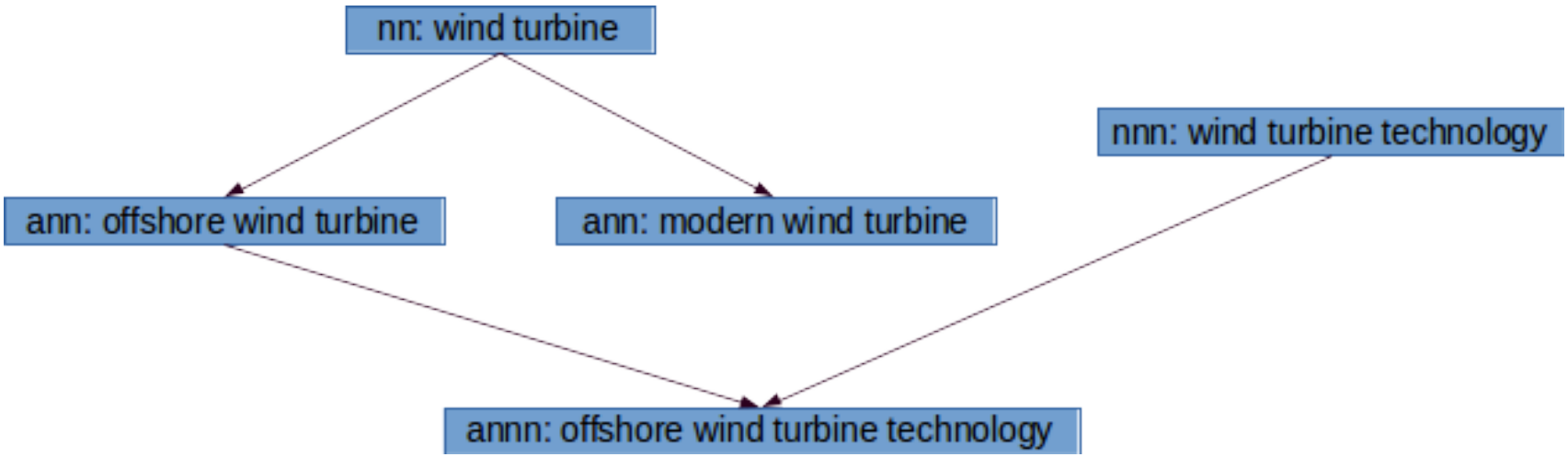


Extraction terminologique : chaîne de traitements

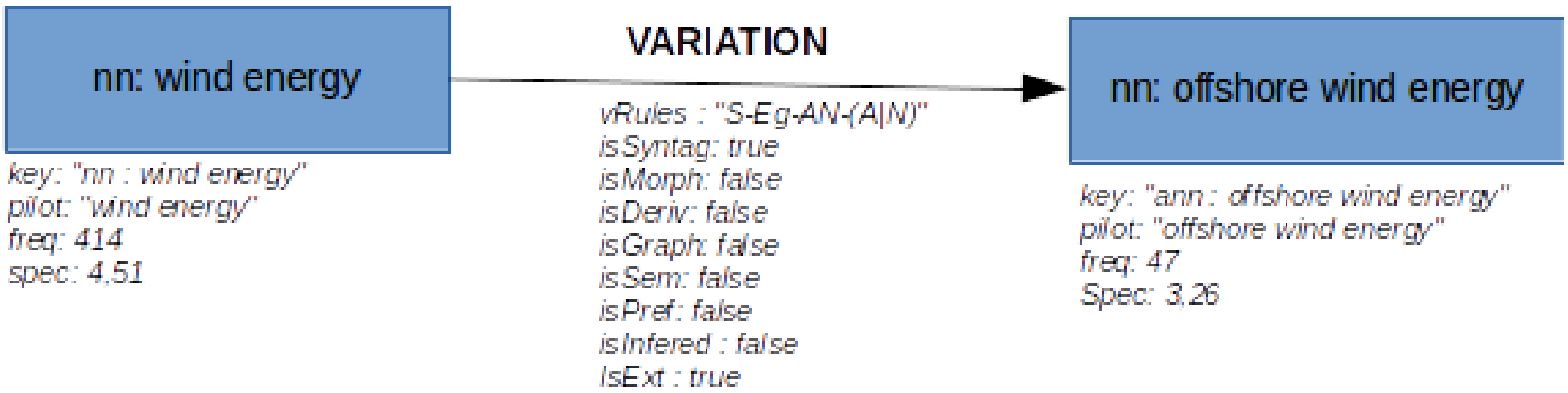


Plan

Modélisation des terminologies à base de graphes (1/2)



Modélisation des terminologies à base de graphes (2/2)



Définition

Variantes sémantiques classiques

- antonymes (synchrone / asynchrone)
- synonymes (coût / prix)
- hyperonymes (électricité / énergie)

Variantes sémantiques de termes complexes

Approche compositionnelle

- $R_1^G : T_1 = T_2 \wedge sem(E_1, E_2) \supset sem(CCT_1, CCT_2)$
- $R_2^G : E_1 = E_2 \wedge sem(T_1, T_2) \supset sem(CCT_1, CCT_2)$

Exemple de règle de variation sémantique

```
"AN-AsynN":
  source: A N
  target: A N
  rule: s[0]==t[0] && synonym(t[1],s[1])
```

Exemple

- low-frequency **noise**
- low-frequency **sound**

Détection des variantes sémantiques

Algorithme

1. la pré-indexation des variantes sémantiques candidates,
2. l'attribution d'un score pour chaque variante sémantique candidate exploitant :
   - la présence dans une ressource sémantique,
   - l'alignement distributionnel (comparaison des contextes),
3. le filtrage des variantes en corpus.

Variation	Dico	Score
power **output** / power **production**	1	0.91
power **output** / power **system**	0	0.45
power **output** / power **producers**	0	0.42

Règles de variations sémantiques

Français

N A → N A'	vents modérés	vents moyens
A N → A' N	générateur synchrone	alternateur synchrone
N P N → N P N'	coût de l'électricité	coût de l'énergie
N P N → N' P N	prix de l'électricité	coût de l'électricité

Anglais

A N → A N'	initial surgery	initial operation
A N → A' N	dynamic stall	static stall
N P N → N P N'	type of surgery	type of operation
N P N → N' P N	center of the blade	midpoint of the blade
N N → N N'	mastectomy swimsuit	mastectomy swimwear
N N → N' N	flow field	exchange field

Performances (1/2)

Temps de traitement

Extraction terminologique	*Durée totale*
Sans variantes sémantiques	33*sec*
Avec variantes sémantiques (dico + distrib)	72*sec*
Avec variantes sémantiques (dico seulement)	55*sec*

`Ubuntu 16.04, Core i7 - 36 documents - 300000 mots`

Performances (2/2)

Évaluation qualitative

Rang	Type	Pilote
2	T	wind turbine
2	V[h]+	wind channel
2	V[h]+	wind farm
2	V[h]+	Enfield-Andreau turbine
3	T	wind energy
3	V[h]+	wind source
3	V[h]+	wind companies
8	T	wind speed
8	V[h]+	Wind turbines-a
8	V[h]+	wind channel
63	T	power production
63	V[h]+	Power Syst
63	V[h]+	power producers
63	V[h]+	**power output**
98	T	energy production
98	V[h]+	electricity production
98	V[h]+	energy consumption
98	V[h]+	**energy yield**
98	V[h]+	energy amount
149	T	tower base
149	V[h]+	**tower foundations**

Export de la terminologie en TSV (1/2)

Extrait

	type	pilot	freq	spec	semScore	isDico	isDistrib
1	T	rotor	848	4,82			
2	T	wind turbine	1855	4,56			
2	V[h]+	wind power-plant	2	1,90	0,97	0	1
2	V[h]+	wind channel	2	2,20	0,97	0	1
2	V[h]+	Wind turbines-a	2	2,20	0,97	0	1
2	V[h]+	wind farm	488	3,20	0,89	0	1
2	V[s]	wind turbine rotor	31	3,38			
2	V[s]+	vertical-axis wind turbine	6	2,37			
2	V[s]	WIND TURBINE APPLICATIONS	86	3,83			
2	V[s]	wind turbine blades	48	3,57			
2	V[h]+	Enfield-Andreau turbine	3	2,37	0,54	0	1
2	V[s]+	wind turbine concepts	37	3,46			
2	V[s]+	wind turbine generator	27	3,02			
2	V[s]+	Domestic Wind Turbines	29	3,35			
2	V[s]+	small wind turbines	33	3,41			
2	V[s]	MW wind turbine	10	2,89			
4	T	wind power	278	4,34			
4	V[s]	wind turbine power	10	2,89			
4	V[s]+	wind power stations	24	3,27			
4	V[s]	Wind Power Project	19	3,17			
4	V[s]	wind power development	14	3,04			
4	V[s]+	wind power generation	11	2,63			
4	V[s]+	wind power capacity	6	2,67			
4	V[s]+	wind power penetration	4	2,50			
4	V[s]	Wind Power Installation	2	2,20			
5	T	airfoil	236	4,26			
6	T	voltage	214	4,22			

Export de la terminologie en TSV (2/2)

Typologie des variations

s	variante syntagmatique
m	variante morphologique : composition
g	variante graphique
h	variante sémantique
d	variante morphologique : dérivation
p	variante morphologique : préfixation
i	variation inférée
V[...]+	la variante possède elle-même des variations

Inférence de variations (1/3)

Problème

Étant donnée une règle morphologique, sémantique, dérivative ou préfixative :

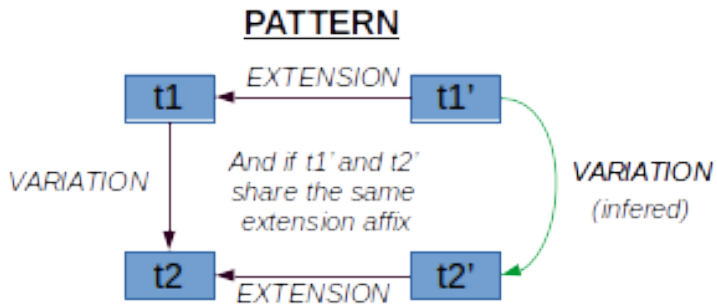
```
"M-S-AN":                    (Ex: horizontal-axis --> horizontal axis)
  source: N [compound]
  target: A N
  rule: s[0][0]==t[0] && s[0][1]==t[1]
```

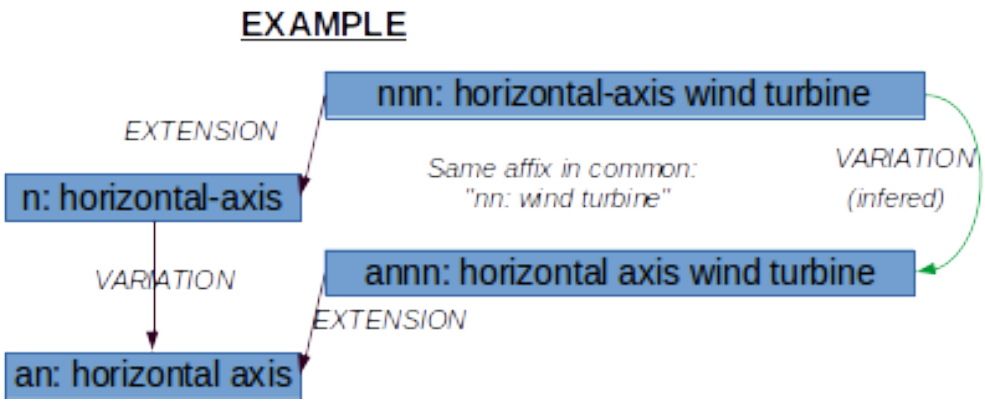
Comment détecter :

horizontal-axis wind turbine → *horizontal axis wind turbine*
offshore horizontal-axis wind turbine → *offshore horizontal axis wind turbine*

Inférence de variations (2/3)

Inférences par recherche de motifs de sous-graphes





Inférence de variations (3/3)

Exemples de variations détectées par inférence

Règle sémantique :

générateur synchrone → alternateur synchrone
générateur synchrone à rotor → alternateur synchrone à rotor

Règle préfixative :

générateur synchrone → générateur asynchrone
générateur synchrone à rotor → générateur asynchrone à rotor

Règle morphologique :

chemotherapy → hormone therapy
individual chemotherapy drugs → individual hormone therapy drugs

Valorisation (1/3)

Conteneur Docker

Lancer TermSuite

Pré-requis : Java 8, TreeTagger

```
java -cp termsuite-core-3.0.4.jar \
       fr.univnantes.termsuite.tools.TerminologyExtractorCLI \
      -t /opt/treetagger \
      -l en \
      --tsv my-termino.tsv \
      -c /path/to/my/corpus
```

Lancer TermSuite avec Docker

Pré-requis : Docker

```
bin/termsuite extract \
      -l en \
      --tsv my-termino.tsv \
      -c /path/to/my/corpus
```

https://github.com/termsuite/termsuite-docker

Valorisation (2/3)

Lanceur Istex

https://github.com/termsuite/termsuite-istex

Lanceur de TermSuite sur l'API ISTEX

```
$ java -cp termsuite-istex-1.1.0.jar \
      fr.univnantes.termsuite.istex.cli.IstexLauncher \
      -t /path/to/tagger \
      -l en \
      --tsv istex-termino.tsv \
      --doc-id F697EDBD85006E482CD1AC91DE9D40F6C629727A,15101397F055B3A872D495F7405D0A3F3E195E0F
```

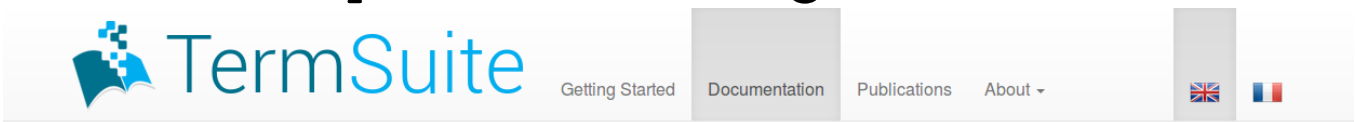
Dockerisé également

https://github.com/termsuite/termsuite-istex-docker

Valorisation (3/3)

Documentation

http://termsuite.github.io/



Documentation TerminologyExtractorCLI

- Getting Started
- Command Line API
  - PreprocessorCLI
  - **TerminologyExtractorCLI**
  - AlignerCLI
- Java API
- Graphical User Interface
- Theory and Architecture
- Resources
- Links

- TerminologyExtractorCLI
  - Usage
  - Description
  - Mandatory options
    - `--from-text-corpus`, `--from-prepared-corpus`
    - `--tsv`, `--tbx`, `--json`
  - Other options
    - `--capped-size` INT
    - `--context-assoc-rate` INT or FLOAT
    - `--context-coocc-th` INT or FLOAT
    - `--context-scope` INT
    - `--contextualize` *(no arg)*
    - `--disable-derivative-splitting` *(no arg)*
    - `--disable-gathering` *(no arg)*
    - `--disable-merging` *(no arg)*
    - `--disable-morphology` *(no arg)*
    - `--disable-native-splitting` *(no arg)*
    - `--disable-post-processing` *(no arg)*
    - `--disable-prefix-splitting` *(no arg)*
    - `--enable-semantic-gathering` *(no arg)*
    - `--encoding`, `-e` ENC
    - `--from-prepared-corpus` DIR

Bilan

Autres améliorations

- post-traitement à base de graphes,
- alignement bilingue,
- embryon d'interface HTTP (`https://github.com/termsuite/termsuite-http`).

Perspectives

- Inférence de variantes par insertion,
- Évaluation des variantes sémantiques,
- Alignement de sous-graphes pour la traduction bilingue,
- Intégration API Istex.

Publications (démos)

Damien Cram and Béatrice Daille. *Terminology Extraction with Term Variant Detection*. Proceedings of ACL-2016 System Demonstrations, 2016.