Grouping Synonyms by Definitions

Ingrid Falk[1], Claire Gardent[2], Evelyne Jacquey[3], Fabienne Venant[4]

[1]INRIA / Université Nancy 2
[2]CNRS / INRIA Nancy Grand-Est, Nancy
[3]CNRS / ATILF, Nancy
[4]Université Nancy 2 / INRIA, Nancy Grand-Est, Nancy

Recent Advances in Natural Language Processing

Outline

Outline

Objective

Build a synonym dictionary which assigns to each meaning of a word the group of synonyms of that word which correspond to this meaning.

A method to merge synonym dictionaries and a large coverage general purpose dictionary.

- 5 synonym dictionaries from the synonym base of the ATILF and
- the TLFi (Trésor de la Langue Française informatisé).

Result :

- A large coverage synonym dictionary with definitions.

Desired results

Example : achever (finish, accomplish)

Synonym dictionaries:

abattre, aboutir, accomplir, aiguiser, améliorer, anéantir, assommer, boucler, cesser, clore, clôturer, complémenter, compléter, conclure, conduire, consommer, continuer, couronner, estoquer, expédier, exécuter, finir, parachever, parfaire, perfectionner, raser, ruiner, réaliser, réussir, se taire, terminer, tuer.

TLFi definitions:

1. *Mettre la dernière main pour perfectionner* (Finalize to improve.)
2. *Porter un coup mortel à un animal déjà atteint physiquement; donner le coup de grâce.* (Give a mortal blow to an animal already physically damaged;)
3. *Mener à sa fin, compléter l'action de.* (Lead to an end, complete the action.)

Desired results

Synonym groupings, attached to the meaning(s) given by the TLFi definitions:

Synonym dictionaries

abattre, aboutir, accomplir, accomplir, aiguiser, améliorer, anéantir, assommer, boucler, cesser, clore, clôturer, complémenter, compléter, conclure, conduire, consommer, continuer, couronner, estoquer, expédier, exécuter, finir, parachever, parfaire, perfectionner, raser, ruiner, réaliser, réussir, se taire, terminer, tuer.

TLFi definitions:

1. Mettre la dernière main pour perfectionner. (Finalize to improve.)
2. Porter un coup mortel à un animal déjà atteint physiquement; donner le coup de grâce. (Give a mortal blow to an animal already achieved physically;)
3. Mener à sa fin, compléter l'action de. (Lead to an end, complete the action.)

Outline

Approach

Input

Synonym dictionaries: Synonym base from the ATILF

General purpose dictionary: TLFi

TLFi $\rightsquigarrow$ definitions $\approx$ meaning (sense)

Output

A synonym dictionary which associates to each sense (definition) of a word the corresponding synonym group.

Related work

- DicoSyn [Manguin et Al., 2004]
- WOLF [Sagot and Fišer, 2008]

Differences

DicoSyn: no synonym groupings, no associated definitions.

WOLF: synonym groupings are obtained by translation to existing BalkaNet synsets,
links to WordNet synsets.

Outline

Resources used

5 of 7 synonym dictionaries from the ATILF.

Syn. Dic.	verbs	Syn/verb
Bailly	2600	1.
Benac	2656	1.5
Du Chazaud	3808	5.25
Larousse	3835	4.7
Le Petit Robert	5027	6.
total	5736	11.

but: no part of speech information, no definitions.

TLFi

- 54 280 entries, 92 997 lemmas, 271 166 definitions.
- digitized, available online (`http://atilf.atilf.fr/`), XML format
- glosses have been lemmatised and POS-tagged.

but: few synonyms, information is not systematic.

Outline

Basic procedure

Given:

- A verb V,
- A set of definitions $D_V^1 \ldots D_V^n$ associated to V by the TLFi,
- The set of synonyms $Syn_V^1 \ldots Syn_V^m$ associated to V by the synonym base.

For each synonym Syn_V^k:

Which are the definitions D_V^i for wich Syn_V^k is synonymous to V?

Basic procedure, ctd.

1. Extract TLFi definitions, convert to index.
   Index = list of content words, lemmatised.
2. Associate indexes to each definition and each synonym.
3. A synonym's index $\rightsquigarrow \cup$ of indexes of each of its definitions.
4. Measure the similarity of two indexes.
   Which definition D_V^i of V is most similar to the definition Syn_V^k?
5. Associate synonyms and definitions. Each synonym is associated to those definitions D_V^i of V which are most similar to the synonym's index.

Extracting the index.

TLFi definitions $\rightsquigarrow$ index

1. We use XML tags to extract TLFi *definitions*.
2. Definitions without a gloss, synonym- or domain indicators are discarded.
3. Index = list of lemmatised content words contained in the gloss, the synonym- and the domain indicators.

Examples

TLFi entry of projeter (to project)

PROJETER, verbe trans.

A. — [Corresp. à *projection*]

1. Jeter loin en avant, avec force. *Il s'agit de projeter l'adversaire sur le tapis (*

terre, dans un heurt un peu violent, il [le cheval] part gaiement (Jeux et sports, 19

• 1. ... elle s'en alla, après m'avoir dévisagé, jugé, pesé, analysé de ce r

eût projeté sur les gens un de ces liquides épais dont se servent les pieu

MAUPASS., *Contes et nouv.*, t.2, Inconnue, 1885, p.1000.

— *Au fig.* Pousser, inciter quelqu'un. V. *jaculation* ex. de Huysmans.

— *Empl. pronom.* Se lancer. *Il s'est projeté dans mes bras!* (CÉLINE, *Mort à crédit,*

2. Envoyer des rayons lumineux; représenter une image sur une surface. *Je vis*

m'élançai (PROUST, *Swann*, 1913, p.35). *Jean Farou (...) projetait sur les murs une*

— *CIN., AUDIOVISUEL.* Passer dans un projecteur (v. ce mot B). *Proje*

cinématographiques de cellules agrandies de telle sorte que leur taille soit

établissements furent obligés d'échanger leurs programmes et, après quelque tem

♦*Projeter quelque lumière* (au fig.). Éclaircir. Synon. *jeter quelque lumière* (v.

Example: the verb *projeter*

Extracted definitions and their indexes:

Definition : Jeter loin en avant avec force. (To throw far ahead and with strength.)

Index : ⟨ jeter, loin, avant, force ⟩

Definition : *CIN. AUDIOVISUEL*. Passer dans un projecteur. (To show on a projector.)

Index : ⟨ cinéma, audiovisuel, passer, projecteur ⟩

Definition : Éclaircir. Synon. jeter quelque lumière. (To lighten, To throw some light.)

Index : ⟨ éclaircir, jeter, lumière ⟩

Measuring the similarity of two indexes.

Experiments with 2 types of similarity measures:

1. Overlap of lemmas (or lemma sequences) between the two indexes.
2. First and second order vector similarity measures with and without TF.IDF cut-off.

Total of 6 similarity measures.

Which similarity measure works best?

Building a reference.

- Gold standard as reference: 27 verbs, their definitions and for each definition the associated synonyms.
- Build triples $\langle$ Verb, Definition, Synonym $\rangle$.
- To a triple $\langle V, D_V, Syn_V \rangle$ we associate
  - the value 1 if Syn_V is considered synonymous to V with the sense given by D_V,
  - the value 0 else.

Example:

$\langle$ achever, Mettre la dernière main pour perfectionner, accomplir $\rangle \rightsquigarrow 1$
($\langle$ perfect, Finalize to improve, accomplish $\rangle \rightsquigarrow 1$)

- Triple $\leftrightarrow$ value associations done by system and annotators
- Result comparisons done on the basis of standard evaluation measures: precision, recall and F-measure.

Building the reference sample.

3 features: genericity, frequency, polysemy

- **genericity** position in EuroWordNet's hierarchy,
- **frequency** extracted from frequency list from 10 years "Le Monde" (newspaper) parsed by Syntex (D. Bourigault).
- **polysemy** number of definitions given by the TLFi.

3 values:

high, medium, low.

$\Rightarrow$ 27 verbs.

4 annotators :

- Fabienne Venant (U. Nancy 2), Mick Grzesitchak (CNRS/ATILF), Christiane Jadelot (CNRS/ATILF), Aurélie Merlot (U. Nancy 2/ATILF).
- Supervised by Evelyne Jacquey (CNRS/ATILF).

Inter-annotator agreement.

Syn. Dic.	Triples	Annotators	Agr.
Robert	2422	P1	80.39
		P2	85.59
		P3	75.55
		Q	59.537
Larousse	2573	P1	79.90
		P2	87.213
		P3	80.10
		Q	63.894
Du Chazaud	4893	P1	81.484
		P2	**87.96**
		P3	73.88
		Q	**57.319**
All dic.	7047	P1	81.481
		P2	87.072
		P3	76.5
		Q	63.374

Outline

Evaluation measures

Recall Number of reference triples identified by the system / Total number of reference triples.

Precision Number of reference triples identified by the system / Total number of triples identified by the system.

F-measure Harmonic mean of recall and precision.

Recall, Precision and F-measure

Sim. measure	**R**	**P**	**F**
baseline	0.497	0.315	0.385
Simple overlap	0.725	0.508	0.598
Extended overlap	0.723	0.508	0.597
Ext. overlap normalised	**0.729**	**0.513**	**0.602**
1^{st} order vectors	0.727	0.510	0.560
2^{nd} order vectors, w/o tf.idf	0.715	0.503	0.590
2^{nd} order, vectors with tf.idf	0.717	0.505	0.592

Table: Precision, recall and F-measure for the various similarity measures. Baseline: random association of synonyms and definitions.

Reflexive usages

- The same TLFi entry may comprise several distinct usages, in particular reflexive and non-reflexive usages – s'abandonner vs. abandonner (abandon oneself vs. abandon).
- These distinct usages often have distinct synonyms:

Example

abandonner	abdiquer, abjurer, abolir, accorder, aliéner, capituler, cesser, concéder, confier, céder, disparaître, donner, délaisser, ...
s'abandonner	céder, faillir, fléchir, mollir, parler, s'adonner, s'effondrer, s'enfoncer, s'ouvrir, s'écouter, satisfaire, se donner, ...

- $\rightsquigarrow$ separate, in a single TLFi entry, definitions of reflexive from non-reflexive usages.
  $\rightsquigarrow$ only compare synonyms of (non-)reflexive usages with definitions of (non-)reflexive usages.

Results with reflexive vs. non-reflexive distinction

	W/o. refl. vs. non-refl. dist.			With refl. vs. non-refl. dist.		
Measure	**R**	**P**	**F**	**R**	**P**	**F**
baseline	0.497	0.315	0.385	0.440	0.433	0.437
Over. 1	0.725	0.508	0.598	0.697	0.685	0.691
Over. 2	0.723	0.508	0.597	0.697	0.685	0.691
Over. 3	**0.729**	**0.513**	**0.602**	**0.711**	0.670	**0.706**
Vect. 1	0.727	0.510	0.560	0.704	**0.693**	0.698
Vect. 2	0.715	0.503	0.590	0.698	0.686	0.692
Vect. 3	0.717	0.505	0.592	0.701	0.689	0.695

Table: Precision, recall and F-measure for various similarity measures with and w/o reflexive vs. non-reflexive distinction.

A more fine-grained linguistic preprocessing improves the results.

Outline

Conclusions

- We presented an unsupervised method to associate synonyms and definitions with reasonably high F-score (0.706 with an inter-annotator agreement of 0.87 as upper bound).
- The linguistic preprocessing has an important impact on the quality of the results.
- The method does not depend on the synonym dictionary, hence several dictionaries may be merged via attachments to TLFi definitions.
- It is applicable to other languages, but results will vary depending on general purpose dictionary used.

Outline

Outlook

- Preprocessing of verbal collocations.
- Applying the method to all synonym dictionaries.
- Extending coverage by integrating further external resources (Wiktionary, Wikipedia, EuroWordNet).
- More comprehensive manual validation through web services.
- Alignment with
  - Wolf (Sagot and Fišer),
  - the French EuroWordNet,
  - the microscopic senses or cliques of the semantic spaces of Dicosyn (work by Ploux, Venant et Al.),
  - Princeton WordNet.

Thank you!

Bibliography

B. Sagot and D. Fišer.
Building a Free French WordNet from Multilingual Resources.
Proc. of Ontolex, 2008.

Jean-Luc Manguin, Jacques François, Rembert Eufe, Ludwig Fesenmeier, Corinne Ozouf and Morgane Sénéchal.
Le dictionnaire électronique des synonymes du CRISCO : un mode d'emploi à trois niveaux.
Les Cahiers du CRISCO, 2004.