Evaluating variants of the Lesk Approach for Disambiguating Words

Florentina Vasilescu
Philippe Langlais
Guy Lapalme

Université de Montréal

Outline

- Fast recap of the Lesk approach (Lesk, 1986)
- Motivations
- Implemented variants
- Evaluation
- Results
- Discussion

The Lesk approach (Lesk, 1986)

Making use of an electronic dictionary

Idea : close-word senses are dependent.

pine	- 1.	kind of **evergreen tree** with needle-shaped leaves ...
	- 2.	waste away through sorrow or illness ...
cone	- 1.	solid body which narrows to a point ...
	- 2.	something of this shape whether solid or hollow ...
	- 3.	fruit of certain **evergreen tree** ...

cone ...**pine** ...?

$$|\text{pine-1} \cap \text{cone-1}| = 0 \quad |\text{pine-2} \cap \text{cone-1}| = 0$$
$$|\text{pine-1} \cap \text{cone-2}| = 0 \quad |\text{pine-2} \cap \text{cone-2}| = 0$$
$$|\text{pine-1} \cap \text{cone-3}| = 2 \quad |\text{pine-2} \cap \text{cone-3}| = 0$$

$\Rightarrow$ **pine-1**

Motivations

Why did we considered the Lesk approach ?

- A simple idea
- An unsupervised method
- A component of some successful systems (Stevenson, 2003)
- Among the best systems at SENSEVAL1... but among the worst at SENSEVAL2 ...
- Some recent promising work (Banerjee and Pedersen, 2003)

Schema of the implemented variants

Input : `t`, a target word

$S = \{s_1, \ldots, s_N\}$ *the set of possible senses, ranked in decreasing order of frequency*

Output : `sense`, the index in S of the selected sense

```
score ← −∞
sens ← 1
C ← Context(t)
for all i ∈ [1,N] do
  D ← Description(s_i)
  sup ← 0
  for all w ∈ C do
    W ← Description(w)
    sup ← sup + Score(D,W)
  end for
  if sup > score then
    score ← sup
    sens ← i
  end if
end for
```

Description of a word

Description(w)

A bag of plain words (nouns, verbs, adjectives and adverbs) in their canonical form (lemma).

1. `Description(w)` $= \bigcup_{s \in Sens(w)}$ `Description(s)`

   with `Description(s)` :

   - DEF plain words of the definition associated to s in WORDNET
     rejection#1 — the act of rejecting something; "his proposals were met with rejection"
     rejection#1 → [act, be, meet, proposal, reject, rejection, something]
   - REL union of the synsets visited while following synonymic and hyperonymic links in WORDNET
     rejection#1 → [rejection, act, human activity, human action]
   - DEF+REL union of DEF and REL

2. `Description(w) = {w}` (simplified variant used by (Kilgarriff and Rosenzweig, 2000))

Context definition

`Context(t)`

1. the set of words centered around the target word `t` : ± 2, ± 3, ± 8, ± 10 et ± 25 words

   - (Audibert, 2003) shown that a symmetrical context is not optimal for disambiguating verbs ($\rightarrow$ $< -2, +4 >$)

   - (Crestan et al., 2003) shown that automatic context selection leads to improvements for some words.

2. words of the lexical chain of `t`

   - term borrowed to (Hirst and St-Onge, 1998)

Context definition

`Context(t)`

lexical chain

Committee approval of Gov._Price_Daniel's "abandoned property" act seemed certain Thursday despite the adamant protests of Texas bankers. Daniel personally led the fight for the measure, which he had watered_down considerably since its rejection by two previous **Legislatures**, in a **public** hearing before the **House_Committee_on_Revenue_and_Taxation**. Under **committee** rules, it went automatically to a **subcommittee** for one week.

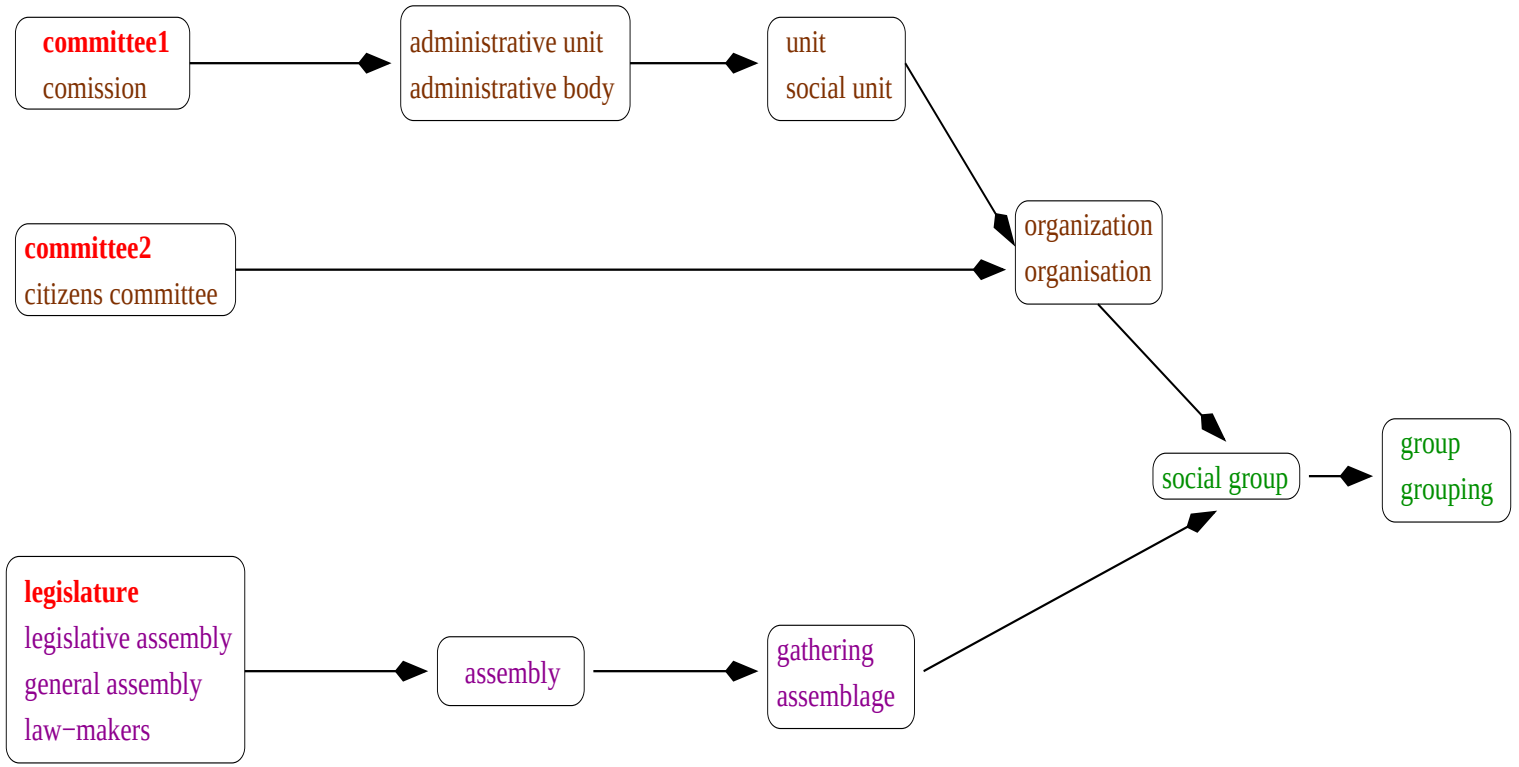
- $E(committee)$ = {committee, commission, citizens, administrative-unit, administrative-body, organization, social-group, group, grouping}
- $E(legislature)$ = {legislature, legislative-assembly, general-assembly, law-makers, assembly, gathering, assemblage, social-group, group, grouping}

$$S(committee, legislature) = \frac{|E(committee) \cap E(legislature)|}{|E(committee) \cup E(legislature)|}$$

Context definition

Context(t)

E1 = {committee, comission, citizens, committe, administrative unit, administrative body, organization, organisation, social group, group, grouping}



E2 = {legislature, legislative assembly, general assembly, law−makers, assembly, gathering, assemblage, social group, group, grouping }

Scoring functions

Score(E_1,E_2)

Cumulative functions of the score given to each intersection between E_1 and E_2.

Lesk each intersection scores 1

Weighted following Lesk's suggestions

- dependence of the size of the entry in the dictionary
- several normalization tested (see (Vasilescu, 2003)), among which the distance between a context-word to the target word

Bayes estimation of $p(s|\text{Context(t)})$, making the naive-based assumption :

$$\log p(s) + \sum_{w \in Context(t)} \log\left(\lambda\, p(w|s) + (1-\lambda)\, p(w)\right)$$

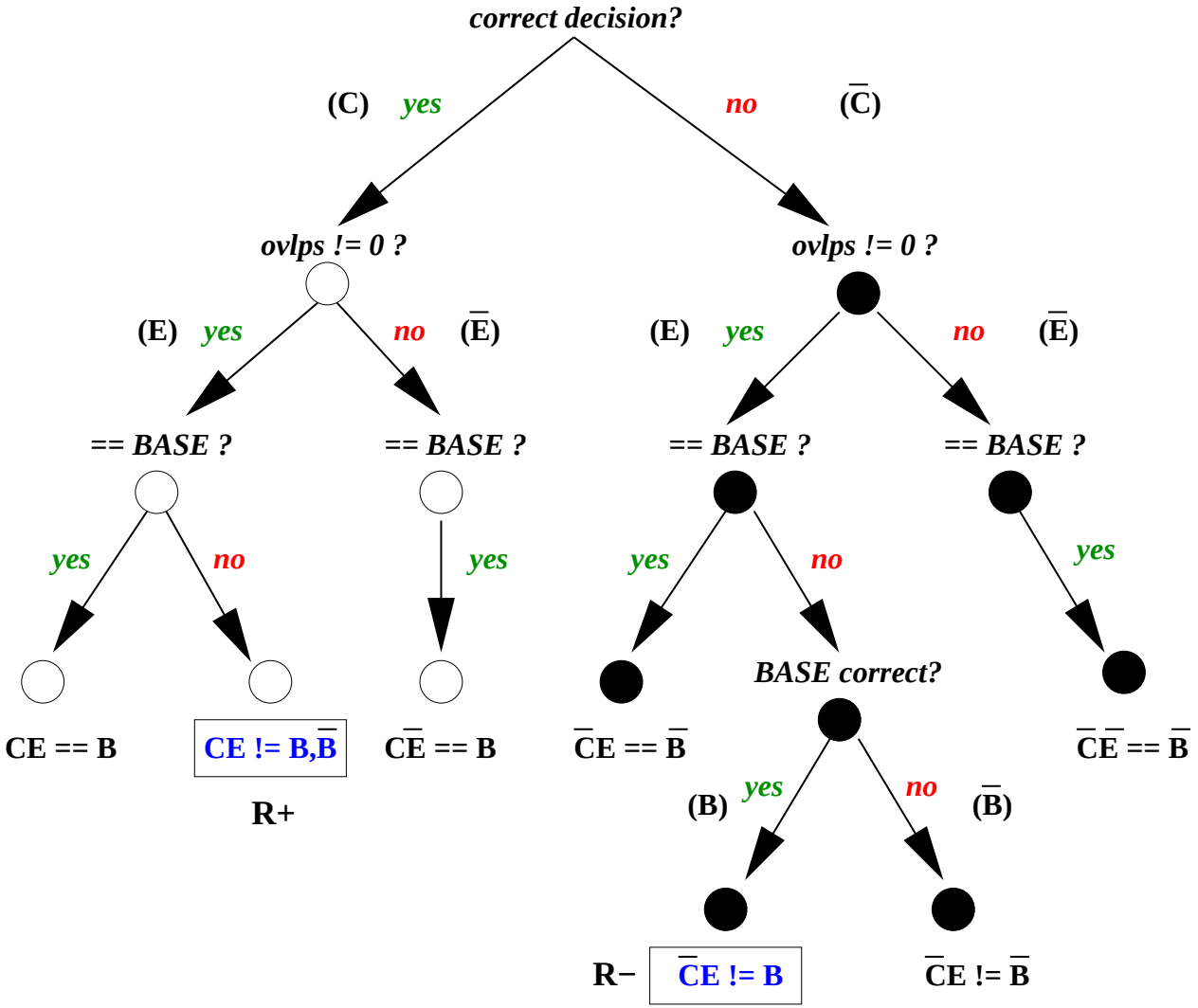
all three distributions $p(s)$, $p(w|s)$ *et* $p(w)$ *"learned" by relative frequency from the* SEMCOR *corpus* ($\lambda = 0.95$ *here)*
$\rightarrow$ supervized method

Protocol

- synsets, definitions and relations taken from WORDNET 1.7.1
- SENSEVAL2 test set, plus several slices of the SEMCOR corpus (cross-validation).
- (task *English all words*)
  $\hookrightarrow$ 2473 target words, over which 0.8% not present in WORDNET
- 2 ways of evaluating the performance
  1. precision & recall rates (SENSEVAL1&2)
  2. risk taken by a variant (according to a taxonomy of decisions a classifier may take)
- 2 baseline systems
  1. most frequent sense (BASE)
  2. BAYES

Evaluation metrics

taxonomy of a decision with respect to a baseline system



Comparing the variants

the DEF variants

	P ±2 R	P ±3 R	P ±8 R	P ±10 R	P ±25 R
LESK	42.6 42.3	42.9 42.6	43.2 42.8	43.3 42.9	42.4 42.0
+ WEIGHTED	39.3 38.9	39.4 39.1	41.2 40.8	40.8 40.4	41.5 41.1
+ LC	58.4 57.9	58.2 57.7	56.2 55.7	55.7 55.2	53.9 53.4

	P ±2 R	P ±3 R	P ±8 R	P ±10 R	P ±25 R
SLESK	58.2 57.7	57.2 56.7	54.7 54.2	53.3 52.8	50.5 50.0
+ WEIGHTED	56.7 56.2	55.5 55.0	51.1 50.6	49.2 48.8	44.4 44.0
+ LC	59.1 58.6	59.1 58.6	58.4 57.9	58.3 57.7	57.4 56.9

	P ±2 R	P ±3 R	P ±8 R	P ±10 R	P ±25 R
BAYES	57.6 57.3	58.0 57.7	56.8 56.6	57.6 57.3	58.5 58.3

BASE : precision of 58 and recall of 57.6

Analyzing the answers

Positive and negative risks

	±2		±3		±8		±10		±25	
	R+	R-	R+	R-	R+	R-	R+	R-	R+	R-
SLesk	3.5	3.3	*3.9*	*4.7*	*6.0*	*9.3*	*6.5*	*11.2*	*7.8*	*15.3*
+ Weighted	*3.5*	*4.8*	*3.9*	*6.4*	*5.9*	*12.8*	*6.4*	*15.2*	*7.8*	*21.3*
+ lc	1.1	0.2	1.2	0.2	1.7	1.3	1.7	1.5	*1.9*	*2.5*

↪ except for lc, the variants take more negative risks than positive, especially for larger contexts

↪ for all the implemented variants, the number of correct answers different from base is very small.

POS filtering

	APOS		RALI		NOPOS	
	P	R	P	R	P	R
SLesk+ lc	61.9	61.3	60.5	59.9	59.1	58.6
base	61.9	61.3	60.4	59.9	57.9	57.6

APOS ≡ the POS is known
RALI ≡ the POS is estimated
NOPOS ≡ the POS is not used

- worth using it . . .
- but does not improve over the base variant when the POS filtering is also applied.

Combining several variants

Oracle simulation

Protocol : the "best" answer is selected among the three best variants selected on a validation corpus.

	SENSEVAL2		SEMCOR	
	F-1	gain%	F-1	gain%
NOPOS				
BASE	57.8	—	66.3	—
oracle	61.0	5.5	70.5	6.2
APOS				
BASE	61.6	—	73.0	—
oracle	68.3	10.9	76.0	4.0

Discussion

- Difficult to improve upon the BASE approach with Lesk variants
- Best approaches tested are those that take less risk (few effective decisions)
- Tendency : performance decreases with larger contexts, best performance observed for 4 to 6 plain-word contexts.
- POS (known or estimated) is worth it (when used as a filter)
- Combining variants might bring clear improvements $\rightarrow$ boosting (Escudero et al., 2000)
- Only local decisions were considered here

Bibliography

L. Audibert. 2003. Étude des critères de désambigïsation sémantique automatique : résultats sur les cooccurrences. In *10e conférence TALN*, pages 35–44, Batz-sur-mer, France, juin.

S. Banerjee and T. Pedersen. 2003. Extended gloss overlaps as a measure of semantic relatedness. In *Eighteenth International Conference on Artificial Intelligence (IJCAI-03)*, pages 805–810, Acapulco, Mexico.

E. Crestan, M. El-Bèze, and C. de Loupy. 2003. Peut-on trouver la taille de contexte optimale en désambiguïsation sémantique ? In *10e conférence TALN*, pages 85–94, Batz-sur-mer, France, juin.

G. Escudero, L. Màrquez, and G. Rigau. 2000. Boosting applied to word sense disambiguation. In *12th European Conference on Machine Learning*, Barceloja, Spain.

G. Hirst and D. St-Onge. 1998. Lexical chains as representation of context for the detection and correction of malapropisms. In C. Fellbaum, editor, *WordNet : An electronic lexical database and some of its applications*, pages 305–331. Cambrige, MA : The MIT Press.

A. Kilgarriff and J. Rosenzweig. 2000. Framework and results for English SENSEVAL. In *Computers and the Humanities*, volume 34, pages 15–48. Kluwer.

M. Lesk. 1986. Automatic sense disambiguation using machine readable dictionaries : How to tell a pine cone from an ice cream cone. In *The Fifth International Conference on Systems Documentation, ACM SIGDOC*.

Mark Stevenson. 2003. *Word Sense Disambiguation. The case for Combinations of Knowledge Sources*. CSLI Studies in Computational Linguistics. CSLI.

F. Vasilescu. 2003. Désambiguïsation de corpus monolingues par des approches de type Lesk. Master's thesis, Université de Montréal.

def or rel ?

Not enough evidence to conclude

SLesk	±2		±3		±8		±10		±25	
DEF	58.2	57.7	57.2	56.7	54.7	54.2	53.3	52.8	50.5	50.0
REL	57.8	57.3	57.5	57.0	56.3	55.8	55.7	55.2	53.0	52.5
DEF+REL	57.3	56.8	56.1	55.6	54.1	53.6	53.0	52.5	50.6	50.1

Most prominent **tendency** : for short contexts (±2), DEF is better. For larger contexts, REL seems more appropriate.