Using WordNet to Build Lexical Sets for Italian Verbs

Anna Feltracco[1,2], Lorenzo Gatti[1,3], Simone Magnolini[1,4],
Bernardo Magnini[1], Elisabetta Jezek[2]

[1]Fondazione Bruno Kessler, [2]University of Pavia, [3]University of Trento, [4]University of Brescia

feltracco@fbk.eu, l.gatti@fbk.eu, magnolini@fbk.eu,
magnini@fbk.eu, jezek@unipv.it

27-30 January 2016, Bucharest

OUTLINE

- What is a lexical set?
- Building lexical sets: goal and motivation
- Methodology: Sentence annotation and lexical sets extraction
  - The Baseline algorithm
  - The LEA algorithm
- Results
- Final considerations and further work

WHAT IS A LEXICAL SET?

Lexical sets are paradigmatic sets of words which occupy the same argument position for a verb, as found in a corpus. (cf. Hanks, 1996 and Jezek and Hanks, 2015)[1]

to read

-> Subject *reads* Object

-> Object *{book, letter, newspaper, report, paper, word, article, story, papers, time, text, mind, page, novel, magazine, poem, passage, ..}* [2]

[1] Hanks P., 1996. Contextual dependencies and lexical sets. *The International Journal of Corpus Linguistics*, 1(1).
Jezek E. and Hanks P., 2010, "What lexical sets tell us about conceptual categories." Lexis 4.7: 22.

[2] Lemmas are extracted from the BNC Corpus, using SketchEngine (Kilgarriff, A. et al., 2004, "Itri-04-08 the sketch engine." Information Technology 105: 116.)

Lexical sets change from verb to verb

- to read – OBJ: *{book, letter, newspaper, report, paper, word, article, story, papers, time, text, mind, page, novel, magazine, poem, passage, bible, ..}*
- to publish – OBJ: *{report, book, article, paper, result, work, letter, study, document,..}*
- to write – OBJ: *{letter, book, article, poem, report, song, name, program, story, word, ..}*
- to send – OBJ: *{letter, message, copy, child, man, troops, money, ..report, .. food,..}*
- to devour – OBJ: *{book, meal, animal, plant, child, Mariana, buffalo, carcass, .. food,.. }*
- to eat – OBJ: *{food, meal meat, fish, breakfast, sandwich, lunch, dinner, bread, diet, ..}*

Lemmas are extracted from the BNC Corpus, using SketchEngine (Kilgarriff, A. et al.,2004, "Itri-04-08 the sketch engine." Information Technology 105: 116.)

Different senses of a verb have different lexical sets

Subject of ‘to rise’ for different senses of the verb:

- to rise up, to rear: *{building, home, church,..}*
- to come up, to uprise: *{sun, moon}*
- to go up, to increase (in value): *{turnover, price, share, rate, unemployment, profit, income, figure, temperature, cost, level, ..}*
- to come up, to move up: *{smoke, ..}*

MOTIVATION

- Verbs' selectional preferences
- Word Sense Disambiguation

  if lexical sets are associated to verb senses -> verb meaning can be induced

Lexical sets for WSD

To rise

- The sun **rose** in the east. → {rise#16, come up#10, uprise#5, ascend#7} [{sun, moon, star}-subj]
- A church **rose** upon that hill. → {rise#4, lift#12, rear#3} [{building, home, church,..}-subj]

MOTIVATION

- Verbs' selectional preferences
- Word Sense Disambiguation

  if lexical sets are associated to verb senses -> verb meaning can be induced
- Semantic Role Labeling -> to automatically annotate roles

Lexical sets for SRL

To rise

- The *land* was silent when the *sun* **rose** in the east.

Rise.01 :

Arg1: *Logical subject, patient, thing rising*

Candidate: "land" and "sun"

[{building, home, church, sun, moon, star}-subj]

no "land" -> Arg1: sun

OUR EXPERIMENT

GOAL: Building lexical sets for argument positions of Italian verbs at sense level

WE NEED:

- a repository of verbs with the specification of their argument structure for each sense
- a repository of sentences associated to each verb sense from which the members of the lexical sets can be extracted

METHODOLOGY

- We use the T-PAS resource [1] , a repository of verb frames for Italian in which :
  - the expected semantic type for each argument slot is specified (e.g. Human, Food, Event, Location, Artifact, …)
  - each frame is related to sentences in a corpus in which the verb is annotated

- In these sentences, we automatically annotate the sets of fillers for the argument slots of the selected verb -> the Baseline Algorithm and the Lea Algorithm

- Both algorithms use a mapping from Semantic types to MultiWordNet synsets [2]

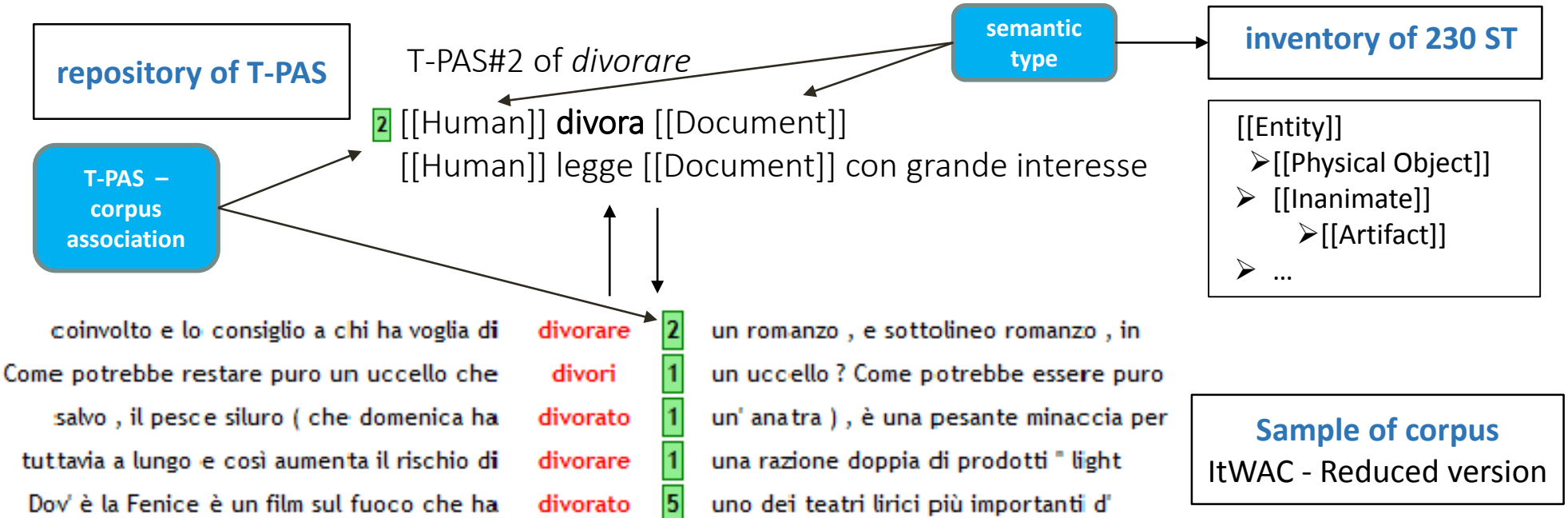
T-PAS resource + MultiWordNet + Sentence Annotation -> Lexical Set

[1] Jezek E. et al., 2014, "T-PAS: a resource of corpus-derived Typed Predicate Argument Structures for linguistic analysis and semantic processing" In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC'14),* Reykjavik, Iceland.

[2] Pianta E. et al., 2002. "MultiWordNet: developing an aligned multilingual database". In *Proceedings of the 1st international conference on global WordNet*, volume 152, pages 55–63.

T-PAS: Typed Predicate Argument Structures

T-PAS is a repository of corpus-derived verb patterns for Italian with specification of the expected semantic type for each argument slot.
T-PASs are acquired following Corpus Patten Analysis methodology (Hanks, 2004).



Visit **tpas.fbk.eu** and download T-PAS

Hanks P., 2004. "Corpus pattern analysis". In *Proceedings of the Eleventh EURALEX International Congress*, Lorient, France, Universite de Bretagne-Sud;

SENTENCE ANNOTATION AND LEXICAL SET BUILDING

Input data from T-PAS

repository of T-PASs

T-PAS#2 of *preparare*
[[Human]] **prepara** [[Food | Drug]]
Eng.: [[Human]] **prepare** [[Food | Drug]]

Sentences

"La nonna, prima di infornare le patate, **prepara** una torta"
Eng. "The grandmother, before baking the potatoes, **prepares** a cake"

For all the sentences = Lexical set

Sentence annotation = annotate lexical items corresponding to Semantic type

[[Human]] – subj = ? [[Food]] – obj = ? [[Drug]] – obj = ?

THE BASELINE ALGORITHM

to identify possible candidate members:

[[Human]] – subj = ? [[Food]] – obj = ? [[Drug]] – obj = ?

1) uses TextPro 2.0[1] for PoS-tagging and lemmatization
2) checks if each lemma is in MWN
3) uses the Semantic type – synsets mapping

Automatic Semantic Type-Synsets mapping

[[Human]] -> human#n

[[Food]] -> food#n

[[Drug]] -> drug#n

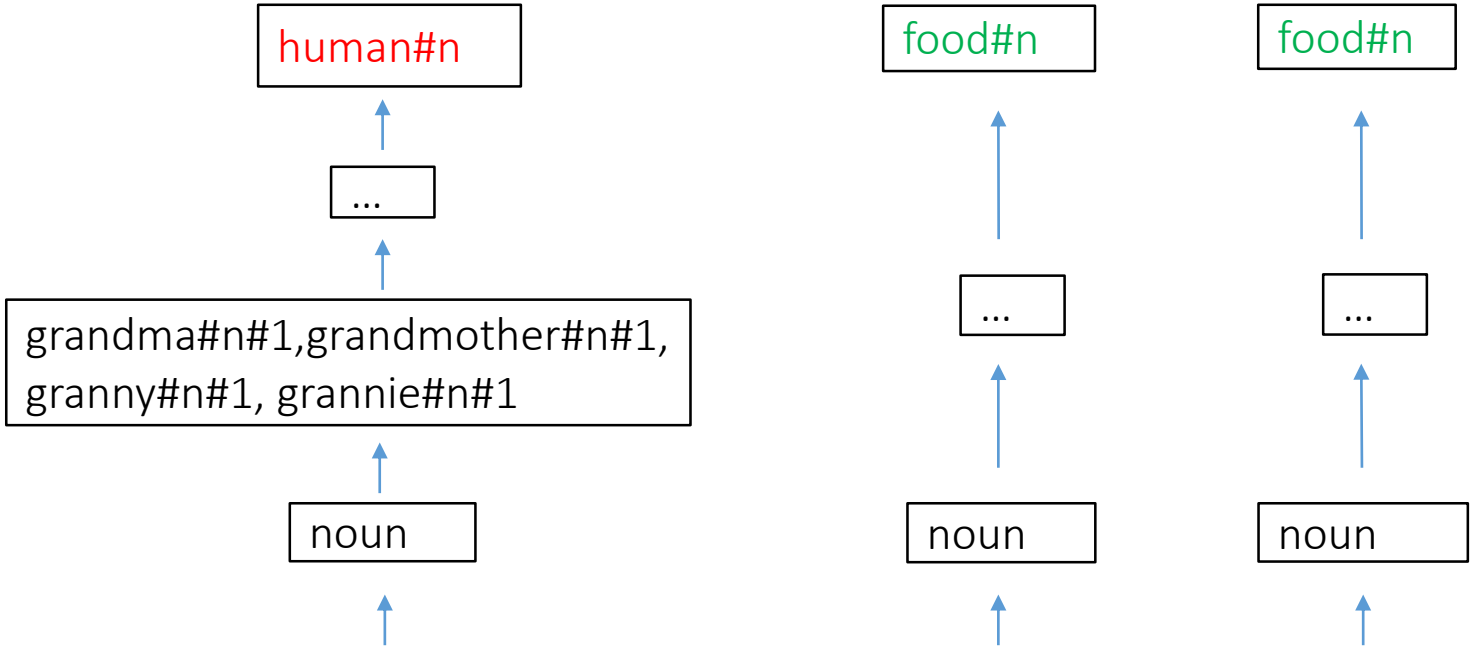
checking if the lemma belongs to a corresponding mapped synset or if it is an hyponym of one such synsets

[1] Pianta E. et al., 2008. The TextPro Tool Suite. In Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC'08), Marrakech, Morocco.

BASELINE:

T-PAS#2 of *preparare*
[[Human]] **prepara** [[Food | Drug]]
Eng.: [[Human]] **prepare** [[Food | Drug]]

[[Human]] – subj = ? [[Food]] – obj = ? [[Drug]] – obj = ?



"La nonna, prima di infornare le patate, **prepara** una torta"
Eng. "the grandmother, before baking the potatoes, **prepares** a cake"

LEA:
THE LEXICAL SET EXTRACTION ALGORITHM

to identify possible candidate members:

[[Human]] – subj = ? [[Food]] – obj = ? [[Drug]] – obj = ?

Baseline +

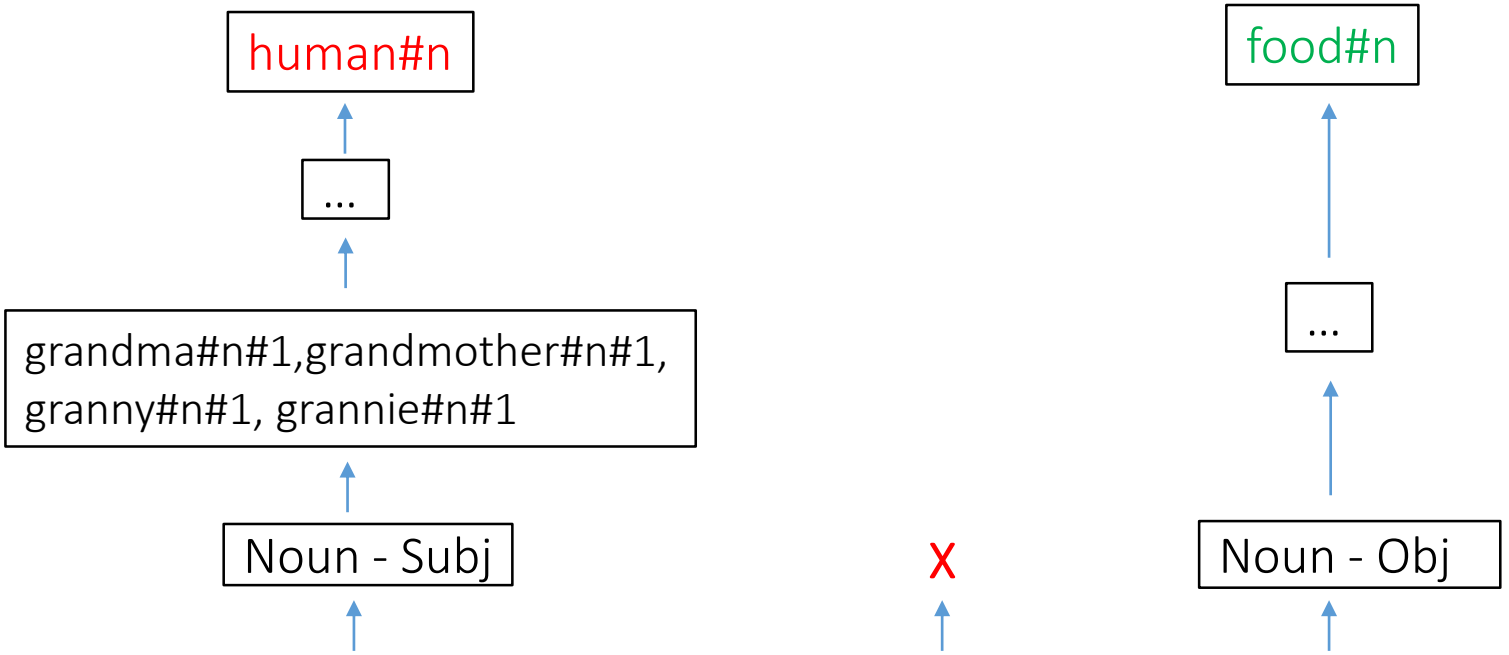
- uses dependency tree of the sentence
- recognizes named entities with TextPro 2.0
- checks for multiword expressions in MWN

-> we expect a higher Precision

LEA: syntactic information

T-PAS#2 of *preparare*
[[Human]] **prepara** [[Food | Drug]]
Eng.: [[Human]] **prepare** [[Food | Drug]]

[[Human]] – subj = ? [[Food]] – obj = ? [[Drug]] – obj = ?

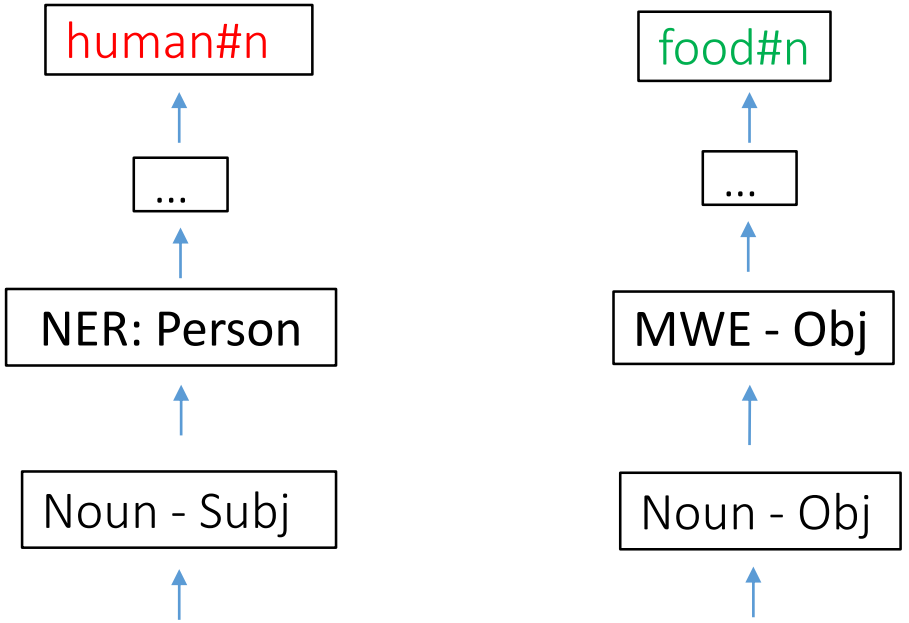


"La nonna, prima di infornare le patate, **prepara** una torta"
Eng. "the grandmother, before baking the potatoes, **prepares** a cake"

LEA: NER and MWE

T-PAS#2 of *preparare*
[[Human]] **prepara** [[Food | Drug]]
Eng.: [[Human]] **prepare** [[Food | Drug]]

[[Human]] – subj = ? [[Food]] – obj = ? [[Drug]] – obj = ?



"Maria Rossi **prepara** la conserva di frutta"
Eng.: "Maria Rossi **prepares** the fruit conserve"

GOLD STANDARD

- 3 annotators manually marked the lexical items or the multiword expressions that correspond to the STs (no pronouns, no relative clauses)

- 500 examples
  (10 sentences x a selection of 10 different STs x 5 different T-PASs;
  e.g. 10 sentences x [[Food]] x 5 T-PASs)

- 981 annotated tokens out of 15090

RESULTS: SENTENCE ANNOTATION

Results for sentence annotation for Baseline and LEA

Automatic mapping			
	Precision	Recall	F1
Baseline	0.28	0.42	0.34
LEA	0.70	0.25	0.37

Evaluation.

Inaccuracies are due to:

- recognition of proper names (Baseline 10 /185 , Lea 26/185)
- PoS tagging step
- dependency parsing step

Results after manual revision of the Semantic Type - synsets mapping

Mapping with manual revision of 11 ST			
Baseline	0.30	0.52	0.38
LEA	0.72	0.32	0.44

- automatic mapping of STs - synsets
- different structure of the two resources (e.g. in T-PAS [[Machine]] is a hypernym of [[Vehicle]], the same is not true for machine#n in MWN)

RESULTS: LEXICAL SET

Similarity between Gold Standard lexical set and lexical set annotated with Baseline and LEA (Dice's coefficient)

5 most populated lexical sets	**Baseline**	**LEA**
Cuocere#2-SBJ-[[Food]] *{pasta, pesce, sugo, carciofo,..}*	0.54	0.57
Crollare#1-SBJ-[[Building]]	0.71	0.60
Dirottare#1-OBJ-[[Vehicle]]	0.83	0.66
Prescrivere#2-OBJ-[[Drug]]	0.42	0.46
Togliere#4-OBJ-[[Garment]]	0.72	0.61

Baseline -> low precision causes major differences with the gold standard sets

LEA -> low recall penalizes the amount of detected items given few sentences to annotate

CONSIDERATIONS AND FURTHER WORK

Final considerations:

- on large scale acquisition, the higher precision for LEA is more promising than the Baseline
- first step on automatic acquisition of lexical sets

Further work:

- extension of the sentence annotation and lexical set population for all T-PAS
- comparison of lexical set in different T-PASs with the same Semantic type

Thank you for your attention