10.12871/CLICIT2014159

Toward Disambiguating Typed Predicate-Argument Structures for Italian

Octavian Popescu, Ngoc Phuoc An Vo, Anna Feltracco, Elisabetta Jezek and Bernardo Magnini
University of Pavia - jezek@unipv.it
FBK, Trento, Italy - magnini|popescu|ngoc|feltracco@fbk.eu

Abstract

English. We report a word sense disambiguation experiment on Italian verbs where both the sense inventory and the training data are derived from T-PAS, a lexical resource of typed predicate-argument structures grounded on corpora. We present a probabilistic model for sense disambiguation that exploits the semantic features associated to each argument position of a verb.

Italiano. *Questo lavoro riporta un esperimento di disambiguazione per verbi italiani, in cui sia la lista dei sensi che i dati di addestramento sono derivati da T-PAS, una risorsa che contiene strutture argomentali tipizzate ricavate da corpora. Presentiamo un modello probabilistico per la disambiguazione che utilizza informazioni semantiche associate a ciascuna posizione argomentale del verbo.*

1 Introduction

Word Sense Disambiguation (WSD) (see (Agirre and Edmonds, 2006) for a comprehensive survey of the topic) is a task in Computational Linguistics where a system has to automatically select the correct sense of a target word in context, given a list of possible senses for it. For instance, given the target word *chair* in the context of the sentence *The cat is on the chair*, and given two possible senses for the word, let's call them *chair_as_forniture* and *chair_as_human*, a WSD system should be able to select the first sense as the appropriate one. An important aspect of WSD is that its complexity is affected by the ambiguity (i.e. the number of senses) of the words to be disambiguated. This has led in the past to discussing various characteristics of available sense repositories (e.g. WordNet, Fellbaum 1998), including the nature and the number of sense distinctions, particularly with respect to the application goals of WSD.

In this paper we address Word Sense Disambiguation of Italian verbs. Differently form previous work on WSD for Italian (Bertagna et al. 2007), where the sense repository was ItalWordNet (Roventini et al. 2003), in our experiments we use verb senses derived from T-PAS, a repository of Typed Predicate Argument Structures for Italian acquired from corpora. There are two benefits of this choice: (i) word sense distinctions are now grounded on actual sense occurrences in corpora, this way ensuring a natural selection with respect of sense granularity; (ii) as in T-PAS for each verb sense a number of sentences are collected, there is no further need to annotate data for training and testing, avoiding the issue of re-interpreting sense distinctions by different people.

The paper is organized as follows. Section 2 introduces T-PAS, including the methodology for its acquisition. Section 3 presents the probabilistic model that we have used for verb disambiguation and Section 4 reports on experimental results.

2 The T-PAS resource

T-PAS (Jezek et al. 2014) is a repository of Typed Predicate Argument Structures (T-PAS) for Italian acquired from corpora by manual clustering of distributional information about Italian verbs, freely available under a Creative Common Attribution 3.0 license [1]. T-PAS are corpus-derived verb patterns with specification of the expected semantic type (ST) for each argument slot, such as [[Human]] guida [[Vehicle]]. T-PAS is the first resource for Italian in which semantic selection properties and sense-in context distinctions of verbal predicates are characterized fully on empirical ground. In the

[1] tpas.fbk.eu.

resource, the acquisition of T-PAS is totally corpus-driven. We discover the most salient verbal patterns using a lexicographic procedure called Corpus Pattern Analysis (CPA, Hanks 2004), which relies on the analysis of co-occurrence statistics of syntactic slots in concrete examples found in corpora.

Important reference points for the T-PAS project are FrameNet (Ruppenhofer et al. 2010), and VerbNet (Kipper-Schuler 2005) and PDEV (Hanks and Pustejovksy 2005), a pattern dictionary of English verbs which is the main product of the CPA procedure applied to English. As for Italian, a complementary project is LexIt (Lenci et al. 2012), a resource providing automatically acquired distributional information about verbs, adjectives and nouns.

T-PAS is being developed at the Dept. of Humanities of the University of Pavia, in collaboration with the Human Language Technology group of Fondazione Bruno Kessler (FBK), Trento, and the technical support of the Faculty of Informatics at Masaryk University in Brno (CZ). The first release contains 1000 analyzed average polysemy verbs, selected on the basis of random extraction of 1000 lemmas out of the total set of fundamental lemmas of Sabatini Coletti 2008, according to the following proportions: 10 % 2-sense verbs, 60 % 3-5-sense verbs, 30 % 6-11-sense verbs.

The resource consists of three components: a repository of corpus-derived T-PAS linked to lexical units (verbs); an inventory of about 230 corpus-derived semantic classes for nouns, relevant for disambiguation of the verb in context; a corpus of sentences that instantiate T-PAS, tagged with lexical unit (verb) and pattern number. The reference corpus is a reduced version of ItWAC (Baroni & Kilgarriff, 2006).

As referenced above, T-PAS specifies the expected semantic type (ST) for each argument slot in the structure; in ST annotation, the analyst employs a shallow list of semantic type labels ([[Human]], [[Artifact]], [[Event]], ecc.) which was obtained by applying the CPA procedure to the analysis of concordances for ca 1500 English and Italian verbs.

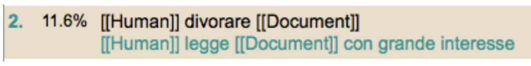
Pattern acquisition and ST tagging involves the following steps:

- choose a target verb and create a sample of 250 concordances in the corpus;

- while browsing the corpus lines, identify the variety of relevant syntagmatic structures corresponding to the minimal contexts where all words are disambiguated;

- identify the typing constraint of each argument slot of the structure by inspecting the lexical set of fillers: such constraints are crucial to distinguish among the different senses of the target verb in context. Each semantic class of fillers corresponds to a category from the inventory the analyst is provided with. If none of the existing ones captures the selectional properties of the predicate, the analyst can propose a new ST or list a lexical set, in case no generalization can be done;

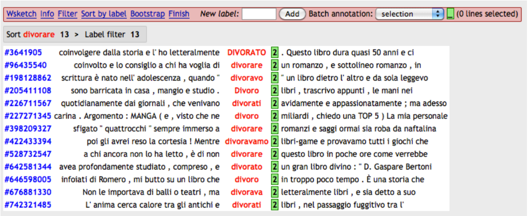
- when the structures and the typing constraints are identified, registration of the patterns in the Resource using the Pattern Editor (see Fig. 1.) Each pattern has a unique identification number, and a description of its sense, expressed in the form of an implicature linked to the typing constrains of the pattern, for example the T-PAS in Fig. 1. has the implicature [[Human]] *legge* [[Document]] *con grande interesse* (read with high interest):



2. 11.6% [[Human]] divorare [[Document]]
[[Human]] legge [[Document]] con grande interesse

Fig 1. Selected pattern for verb *divorare*

- assignment of the 250 instances of the sample to the corresponding patterns, as shown in Fig. 2:



```
#3641905     coinvolgere dalla storia e l' ho letteralmente  DIVORATO   2 . Questo libro dura quasi 50 anni e ci
#96435540         coinvolto e lo consiglio a chi ha voglia di  divorare   2 un romanzo , e sottolineo romanzo , in
#198128862   scrittura è nato nell' adolescenza , quando "     divoravo   2 " un libro dietro l' altro e da sola leggevo
#205411108       sono barricata in casa , mangio e studio .    Divoro     2 libri , trascrivo appunti , le mani nei
#226711567   quotidianamente dai giornali , che venivano       divorati   2 avidamente e appassionatamente ; ma adesso
#227271345 carina . Argomento : MANGA ( e , visto che ne       divoro     2 miliardi , chiedo una TOP 5 ) La mia personale
#398209327        sfigato " quattrocchi " sempre immerso a     divorare   2 romanzi e saggi ormai sia roba da naftalina
#422433394           poi gli avrei reso la cortesia ! Mentre   divoravamo 2 libri-game e provavamo tutti i giochi che
#528732547            a chi ancora non lo ha letto , è di non  divorare   2 questo libro in poche ore come verrebbe
#642581344   avea profondamente studiato , compreso , e        divorato   2 un gran libro divino : " D. Gaspare Bertoni
#646598005   infoiati di Romero , mi butto su un libro che     divoro     2 in troppo poco tempo . È una storia che
#676881330          Non le importava di balli o teatri , ma    divorava   2 letteralmente libri , e sia detto a suo
#742321485         L' anima cerca calore tra gli antichi e     divorati   2 libri , nel passaggio fuggitivo tra l'
```

Fig 2. Sample annotation for pattern 2 of *divorare (devour)* - SketchEngine

In this phase, the analyst annotates the corpus line by assigning it the same number associated with the pattern.

3 Disambiguation Method

In this section, we present a disambiguation method for corpus patterns and apply it to the task of verb disambiguation with respect to the T-PAS resource. The method is based on identifying the important elements of a pattern which are disambiguating the verb in the text. The importance of such elements is evaluated according to their effect on the sense of the verb, expressed as a relationship between the senses of the words inside a pattern. It has been noted that the relationship between verb meaning and semantic types is constrained, such that the context matched by a pattern is the sufficiently minimal context for disambiguation. This relationship, called chain clarifying relationship (CCR), is instrumental in doing pattern matching as well as in finding new patterns. (Popescu & Magnini 2007, Popescu 2012).

From a practical point of view, the probability of occurrence of a word and the probability of the verb are independent given the probability of the semantic type. As such, the CCR is very efficient in dealing with sparseness problem. This observation has a big positive impact on the disambiguation system, because it directly addresses two issues:1) the necessity of large quantity of training alleviating the data sparseness problem (Popescu 2007, Popescu 2013) and 2) the overfitting of probabilities, with important consequences for the disambiguation of less frequent cases (Popescu et. al 2007). The method divides the vocabulary in congruence classes generated by CCR for each verb and we build a classifier accordingly (Popescu 2013 and Popescu et al. 2014). To this end, we carry out an automatic analysis of the patterns at the training phase, which allows us to compute a confusion matrix for each verb pattern number and congruence class. The exact procedure is presented below.

We introduce here a probabilistic system which does partial pattern matching in text on the basis of individual probabilities which can be learned form training. Matching a corpus pattern against a verbal phrase involves labelling the heads of the constituents with semantic features and the verb with a pattern number. We build a probabilistic model in which we compute the probability in Equation (1).

$$p(t_0, t_1, t_2, t_3, .., t_n, w_1, w_2, w_n) \qquad (1)$$

where t_0 is the pattern number, t_i is the semantic type of the word w_i, which is the head of the ith constituent, with i from 1 to n. For a given sentence we choose the most probable labeling, Equation (2)

$$p(t^c_0, t^c_1, t^c_2, t^c_3, .., t^c_n, w^c_1, w^c_2, w^c_n) = \arg\max_{t_i} p(t_0, \mathrm{w}_n) \qquad (2)$$

On the basis of the relationship existing between the senses of the fillers of the corpus pattern given by CCR, and the fact that the patterns have a regular language structure, we learn for each verb its discriminative patterns with semantic types. Using the chain formula, and grouping the terms conveniently, Equation (1) becomes Equation (3).

$$\begin{aligned} p(t_0, t_1, t_2, t_3, .., t_n, w_1, w_2, w_n) \ &= p(t_0)p(w_1|t_0)... \\ &... p(t_n|t_0, w_1, t_1, w_2,..., t_{n-1}, w_n) \\ &\simeq \mathrm{p(t_0)}p(w_1|t_0)pt_1|t_0, w_1)p(w_2|t_0)p(t_2|t_0, w_2) \ ... \\ &... p(t_n|t_0, w_n) \\ &\simeq \mathrm{p(t_0)}p(w_1|t_0)p(t_1|t_0)p(t_1|w_1)p(w_2|t_0)p(t_2|t_0)p(t_2|w_2) \ ... \\ &... p(t_n|t_0)p(t_n|t_0)p(t_n|w_n) \end{aligned} \qquad (3)$$

The quantities on the right hand side are computed as follows:

- $p(t_0)$ is the probability of a certain pattern. This probability is estimated from the training corpus, via ratio of frequencies.
- $p(w_i|t_0)$ is the probability of a certain word to be the head of a constituent. We used the Italian dependency parser MaltParser (Lavelli et al. 2009) for determining the head of the syntactic constituents and their dependency relationships. However, we allow for any content word to fulfill the role of subject, object or prepositional object with a certain probability. This probability is set a priori on the basis of the parser's probability error and the distance between the word and the verb.
- $p(t_i|t_0, w_i)$ is the probability that a certain word at a certain position in the pattern carries a specific semantic type. This probability is equated to $p(t_i|w_i)p(t_i|t_0)$, assuming independence between the verb sense and the word given the semantic type. The first of the two later probabilities is extracted from Semcor (Miller et al. 1993, Pianta el al. 2002) and Lin distance (Lin 1998), considering the minimal distance between a word and a semantic type. The second probability is computed at the training phase considering the frequency of a semantic type inside the pattern.

The probabilities may be affected by the way the training corpus is compiled. It is assumed that the examples have been drawn randomly from a large, domain independent corpus. We call the resulting model the CF_CCR, from chain formula with CCR.

4 Experiments and Results

We performed the following experiment: we have considered all the verbs present in T-PAS at this

System	*Attribute*	*Macro Average*
5libSVM	5 words	67.871
10libSVM	10 words	65.556
CF_CCR	syn-sem	71.497

Table 1: Direct global evaluation

moment. We have excluded the mono pattern verbs, as in this case there is no ambiguity, and we are left with 916 verbs. For each verb we split the T-PAS examples into train and test considering 80% and 20% respectively. We trained two SVM bag of words models considering a window of 5 and 10 tokens around the target verb. We used the WEKA libsvm library and we compared the results against the model presented in the previous section. The global results, macro average, are presented in Table 1. We report the precision here, corresponding to the true positive, as we annotated all examples.

The model 10libSVM performed worse than the other two, probably due to the noise introduced. On the other hand, it is surprising how well the 5 window model performed. We reckon that this is because of the fact that most of the time the direct object is within 5 words distance from the verb and the majority of the T-PAS patterns consider the direct object as the main distinctive feature, and the set of words occurring in the T-PAS examples is relatively small. Therefore the probability of seeing the same word in test and train is big.

We considered to investigate more the distribution of the results. For this, we considered the better performing bag of word system, namely five words window system, 5libSVM against CF_CCR.

The variation of precision is actually large. It ranges from below 10% to 90%.The number of verbs which are disambiguated with a precision bigger than 60% represents the large majority with 72% of the verbs. This suggests, that instead of macro average, a more indicative analysis could be carried out by distinguishing between precision on verbs with different number of patterns.

We looked into the influence of the number of patterns for the accuracy of the systems. As expected, the bigger the number of patterns the less precise is the system. The extremities are the ones that have an accelerated rate of growth. For example, the precision over 90% and under 10% goes from the biggest (lowest) coverage for 2 patterns, to lowest (bigger) for 9 patterns. The behaviour of CF_CCR is somewhat different from 5libSVM, the CF_CCR is able to identify more senses, thus achieves a better precision for verbs with more than 6 patterns, than 5libSVM does. In Table 2 we present comparatively how many times the system makes a less than 50% accurate prediction for patterns that have a frequency greater than 10%. As we can see, the CF_CCR system is between 6% to 20% percent better than 5libSVM in coping with these cases, proving that combining syntactic and semantic information reduces the overfitting. The fact that the absolute number decreased also with the number of patterns is due mainly to the fact that also the number of examples decreases drastically.

No. Patterns	*5libSVM*	*CF_CCR*
2	57	53
3	118	109
4	126	114
5	112	98
6	85	77
7	50	44
8	29	23
9	28	21

Table 2: Errors on patterns with frequency >10%

5 Conclusion

We have presented a word sense disambiguation system for Italian verbs, whose senses have been derived from T-PAS, a lexical resource that we have recently developed. This is the first work (we hope that many others will follow) attempting to use T-PAS for a NLP task. The WSD system takes advantage of the T-PAS structure, particularly the presence of semantic types for each verbal argument position. Results, although preliminary, show a very good precision.

As for the future, we intend to consolidate the disambiguation methodology and we aim at a more detailed annotation of the sentence argument, corresponding to the internal structure of verb patterns. We plan to extend the analysis of the relationship between senses of the different positions in a pattern in order to implement a metrics based on tree and also to substitute the role of the parser with an independent pattern matching system. The probabilistic model presented in Section 3 can be extended in order to determine also the probability that a certain word is a syntactic head.

Acknowledgments

This work was partially supported by the EC-funded project EXCITEMENT (FP7ICT-287923) and by the BCROCE project of the autonomous province of Trento.

Reference

E. Agirre and P. Edmonds, 2006. *Word sense disambiguation: algorithms and applications*, Springer.

M. Baroni and A. Kilgarriff. 2006. Large Linguistically-Processed Web Corpora for Multiple Languages. In *EACL 2006 Proceedings*, pp. 87-90.

F. Bertagna, A. Toral and N. Calzolari, 2007. *EVALITA 2007: THE ALL-WORDS WSD TASK*, in *Proceedings of Evalita 2007*, Rome.

C. Fellbaum, 1998. *WordNet: An Electronic Lexical Database*, Cambridge, MA: MIT Press.

P. Hanks. 2004. Corpus Pattern Analysis. In G. Williams and S. Vessier (eds.). *Proceedings of the XI EURALEX International Congress*. Lorient, France (July 6-10, 2004), pp. 87-98.

P. Hanks and J. Pustejovsky. 2005. A Pattern Dictionary for Natural Language Processing. In *Revue française de linguistique appliquée*, 10 (2).

E. Jezek, B. Magnini, A. Feltracco, A. Bianchini and O. Popescu. 2014. T-PAS: A resource of Typed Predicate Argument Structures for linguistic analysis and semantic processing. In N. Calzolari et al. (eds) *Proceedings of the Ninth International Conference on Language Resources and Evaluation* (LREC'14), Reykjavik, Iceland (May 26-31, 2014), Paris: European Language Resources Association (ELRA), 890-895.

K. Kipper-Schuler. 2005. *VerbNet: A broad coverage, comprehensive verb lexicon*. Ph.D. Thesis. University of Pennsylvania, USA.

A. Lenci, G. Lapesa and G. Bonansinga. 2012. LexIt: A Computational Resource on Italian Argument Structure. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation* (LREC12), Istanbul (May 23-25, 2012), pp. 3712-3718.

A. Lavelli, J. Hall, J. Nilsson and J. Nivre. 2009. MaltParser at the EVALITA 2009 Dependency Parsing Task. In *Proceedings of EVALITA 2009 Workshop on Evaluation of NLP Tools for Italian*, Reggio Emilia, Italy.

E. Pianta, L. Bentivogli and C. Girardi. 2002. MultiWordNet: Developing an aligned multilingual database. In *Proceedings of the 1st International Global WordNet Conference*, Mysore

D. Lin. 1998. An information-theoretic definition of similarity. In *Proceedings of the Fifteenth International Conference on Machine Learning (ICML)*, Madison, Wisconsin, USA.

O. Popescu and B. Magnini. 2007. Sense Discriminative Patterns for Word Sense Disambiguation. In *Proceedings of the SCAR Workshop 2007, NODALIDA*, Tartu.

O. Popescu, S. Tonelli, Sara and E. Pianta. 2007. IRST-BP: Preposition Disambiguation based on Chain Clarifying Relationships Contexts. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, Prague.

O. Popescu. Building a Resource of Patterns Using Semantic Types. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation* (LREC-12), Istanbul.

O. Popescu. 2013. Learning Corpus Pattern with Finite State Automata. In *Proceedings of the ICSC 2013*, Postadam.

O. Popescu, M. Palmer, P. Hanks.2014. In *Mapping CPA Patterns onto OntoNotes Senses*. LREC 2014: 882-889.

A. Roventini, A. Alonge, F. Bertagna, N. Calzolari, J. Cancila, C. Girardi, B. Magnini, R. Marinelli, M. Speranza, A. Zampolli, 2003. *ItalWordnet: Buiding a Large Semantic Database for the Automatic Treatment of Italian.* In *Zampolli A., Calzolari N., Cignoni L. (eds.), Computational Linguistics in Pisa (Linguistica Computazionale a Pisa), Linguistica Computazionale, Special Issue, Vol. XVI-XIX, pag. 745-791.*

J. Ruppenhofer, M. Ellsworth, M. R. L. Petruck, C. R. Johnson and J. Scheffczyk. 2010. *FrameNet II: Extended theory and practice*. International Computer Science Institute, University of Berkeley. (Manuscript, Version of September 14, 2010).

F. Sabatini and V. Coletti 2007. *Il Sabatini-Coletti. Dizionario della lingua italiana 2008*, Milano, Rizzoli-Larousse.