Enriching the "Senso Comune" Platform with Automatically Acquired Data

Tommaso Caselli, Laure Vieu, Carlo Strapparava, Guido Vetere
TrentoRISE, IRIT-CNRS & LOA-CNR, FBK, IBM-CAS Trento
Via Sommarive, 18 Povo I-38123, 118 route de Narbonne Toulouse F-31062, Piazza Manci, 17 Povo I-38123
t.caselli@trentorise.eu, vieu@irit.fr, strappa@fbk.eu, gvetere@it.ibm.com

Abstract

This paper reports on research activities on automatic methods for the enrichment of the Senso Comune platform. At this stage of development, we will report on two tasks, namely word sense alignment with MultiWordNet and automatic acquisition of Verb Shallow Frames from sense annotated data in the MultiSemCor corpus. The results obtained are satisfying. We achieved a final F-measure of 0.64 for noun sense alignment and a F-measure of 0.47 for verb sense alignment, and an accuracy of 68% on the acquisition of Verb Shallow Frames.

Keywords: Word Sense Alignment, Verb Frame Acquisition, Lexico-semantic Resource

1. Introduction

This paper describes current research activities on automatic methods for the enrichment of the Senso Comune[1] platform to achieve a robust and interoperable lexico-semantic resource for Italian. At this stage of development, we will report on two tasks, namely word sense alignment and automatic extraction of verb shallow frames, as a way for achieving conceptual interoperability (Witt et al., 2009) (Fang, 2012) among different language resources.

The remainder of this paper is organized as follows: Section 2. will shortly presents the Senso Comune Initiative and its model. Section 3. is focused on two case studies for Word Sense Alignment (WSA, henceforth) as a preliminary and necessary step to make Senso Comune conceptually interoperable with other lexico-semantic resources. We focused on verb and noun alignment between the Senso Comune Lexicon and the Italian section of MultiWordNet (Pianta et al., 2002). In Section 4. we will report on the automatic acquisition of Verb Shallow Frames (VSFs, henceforth) as a strategy to achieve interoperability on other levels of linguistic analysis. VSFs have been extracted from the Italian section of the MultiSemCor corpus (Bentivogli and Pianta, 2005) and compared with automatically acquired VSFs from a large corpus of Italian (the La Repubblica Corpus, (Baroni et al., 2004)). The data thus collected, which associate specific VSFs to verb senses (namely synsets) will provide a basis for the development of a VerbNet-like lexicon for Italian. Finally, section 5. will draw on conclusions and future work.

2. The Senso Comune Initiative: a short introduction

Senso Comune (SC) aims at building an open knowledge base for the Italian language, designed as a crowd-sourced initiative that stands on the solid ground of an ontological formalization and well established lexical resources. The SC platform is specified in three modules comprising a *top level module*, which contains basic ontological concepts and relations, inspired by DOLCE (Descriptive Ontology for Linguistic and Cognitive Engineering) (Masolo et al., 2002), a *lexical module*, which models general linguistic and lexicographic structures, and a *frame module* providing concepts and axioms for modeling the predicative structure of verbs, nouns and adjectives. On the top level module, DOLCE basic ontological distinctions are kept. For instance, DOLCE's Endurant and Perdurant match Senso Comunes Continuant and Occurrent, respectively. The main difference with respect to DOLCE top level is represented by the merging of DOLCEs Abstract (e.g. mathematical entities, dimensional regions, ideas) and Non-Physical Endurant (e.g. social objects) categories into the Senso Comune category Non-Physical Entity.

The adoption of a legacy dictionary as a foundation for the resource has led to modeling SC on a distinction between lexicographic structures and linguistic facts. Similarly to models like LMF (Buitelaar et al., 2009) and Lemon (Chiarcos et al., 2013), the purpose is to provide a structure to accommodate linguistic resources where lexical units are associated with their acceptations. In SC, the distinction between the lexicographic meanings and relationships (such as synonymy, hyponymy, antinomy, among others) from the formal account of their phenomenal counterparts (e.g. concepts, equivalence, inclusion, disjointness) brings a number of benefits. As an immediate consequence, this separation prevents that lexicographic entries will be directly mapped to logic propositions, while preserving the possibility of relating entries from a lexicon to any suitable ontology. For instance, SC separates the notion of lemma from that of lexeme. A lemma in SC captures the section of a dictionary where an etymologically consistent bundle of senses (called Meaning Record) of a given lexeme is described by means of a lexicographic apparatus (e.g. definition, grammatical constraints, usage examples). A Meaning Record is part of a Lemma whose instances and attributes form the body of the SC lexicon. Each instance of a Meaning Record, where a specific sense of a Lemma is described, can be mapped to the Meaning class. Such a mapping between instances of Meaning Record and the Meaning class can be done by exploiting different mechanisms in OWL2 syntax (e.g. punning, annotations, among others). An important aspect is that different Meaning Record instances can be mapped to the same Meaning class, thus facilitat-

[1] http://www.sensocomune.it/

ing the mappings of meaning record instances from different language resources (e.g. such as for the alignment between the MultiWordNet synsets and the SC lexicon entries). Similarly, lexical relations do not have any ontological import and correspondences between lexical relations and ontological relations must be introduced on the basis of dedicated heuristics[2].

The lexicon entries have been obtained from a reverse engineering procedure from the De Mauro GRADIT dictionary (De Mauro, 2000) and consists of 2,071 fundamental Italian lexemes[3] for a total of 11,939 meanings. Verbs account for 3,827 senses, corresponding to 643 lemmas, with an average polysemy of 5.9 senses per lemma, while nouns have 4,586 senses, corresponding to 1,111 lemmas with an average polysemy of 4.12 senses per lemma. All nominal entries have been manually classified according to the ontological classes described in the SC ontological module. A classification of the verb entries will start in the near future. Currently in SC, word senses are not hierarchically structured and no semantic relation is extensively encoded (so far, synonyms relations for noun senses amount to 49 entries, covering less the 5% of the fundamental noun senses). This means that with respect to other lexico-semantic resources, such as WordNet, senses of polysemous entries have a flat representation, where each (fundamental) meaning is report one following the other.

3. Conceptual Interoperability with Word Sense Alignment

Following (Matuschek and Gurevych, 2013), WSA can be formally defined as a list of pairs of senses from two lexical-semantic resources. A pair of aligned senses denotes the same meaning. To clarify, consider this example for two sense descriptions of the word "*day*", taken from translated SC Lexicon and MultiWordNet (MWN), respectively:

- `amount of hours of work done in one day` [SC Lexicon]
- `the recurring hours established by contract or usage for work` [MWN]

The two sense descriptions are equivalent and refer to the same meaning, thus they must be aligned.

MWN is a computational multilingual lexicon perfectly aligned to Princeton WN 1.6. As in WN, concepts are organized in synonym sets (synsets), hierarchically connected by means of hypernym relations, and includes other semantic relations such as parthood and troponymy among others[4]. The main motivations for the selection of MWN for the WSA can be summarized as follows: i.) WN/MWN is one of the most used lexico-semantic resource for different NLP tasks; ii.) WN/MWN has already been used as a pivot lexicon both for the alignment and for the merging of other lexica or language resources (for instance, (Niemann and Gurevych, 2011); (Navigli and Ponzetto, 2012) for the alignment of WN and Wikipedia; (Navigli, 2006) for the alignment of WN and the Oxford English Dictionary; (Shi and Mihalcea, 2005) for the integration of WN, VerbNet and FrameNet), thus opening up the possibility of connecting and making interoperable SC with others lexico-semantic resources, not only in Italian; iii.) WN/MWN can provide a taxonomic structure to the entries of the SC lexical module. At the same time, WN/MWN can also benefit from the alignment with the SC lexicon. In particular, i.) the introduction of high quality glosses in Italian: only 3,177 synsets (8,21%) over a total of 38,653 composing the Italian section of MWN are in Italian, while the remaining have inherited the original English gloss in WN 1.6; ii.) the assignment of a foundational ontological class to each synset to facilitate the identification of taxonomic errors or suggest better sense descriptions; and iii.) an improvement in the coverage of the MWN entries for Italian.

Although in previous works on WSA different methods are employed (similarity-based approaches vs. graph-based approaches), they have common elements such as: i.) the extensive use of lexical knowledge based on the sense descriptions; e.g.: WN glosses; an article first paragraph as in the case of Wikipedia; and ii.) the extension of the basic sense descriptions with additional information; e.g.: hypernyms for WN entries, domains labels or categories for dictionaries or Wikipedia entries. With respect to other resources which have been sense aligned, the SC Lexicon has some shortcomings, namely i.) no distinction between core senses and subsenses for polysemous entries; ii.) no presence of hypernyms or taxonomic structures in the entries; and iii.) no domain labels (e.g. *Biology*, *Architecture*, *Sport* ...) associated with senses. Moreover, the low number of MWN glosses in Italian prevents a straightforward application of state-of-the-art methods for sense alignment. MWN sense descriptions must be built up from other sources. The main issue we are facing is related to data sparseness, that is how to tackle sense alignment when we have few descriptions in Italian (MWN side) and few meta-data and no structuration over senses (SC side).

We focused our alignment tasks on verb and noun senses in order to understand the feasibility of this task by means of automatic approaches. Provided the limits both of the MWN Italian section and of the SC Lexicon, we decided to apply two methods, namely Lexical Overlap and Sense Similarity.

3.1. Lexical Overlap

In the Lexical Overlap method, for each word w and for each sense s in the given resources $R \in \{\text{MWN, SC}\}$ we constructed a sense descriptions $d_R(\text{s})$ as a bag of words in Italian. Provided the different characteristics of the two resources, two different types of bag of words have been built. As for the SC Lexicon, the bag of words is represented by the lexical items in the textual definition of s_w, automatically lemmatized and part-of-speech analyzed with the TextPro tool suite (Pianta et al., 2008) with standard stopword removal. On the other hand, for each synset, *S*, and for each part of speech in analysis, the sense description of each MWN synset was built by optionally exploiting:

- the set of synset words in a synset excluding w;

[2]For additional details on the Senso Comune model see details see (Vetere et al., 2012), (Oltramari et al., 2013).

[3]Lexemes covering about 90% of all spoken and written texts in Italian. See (Oltramari et al., 2013).

[4]Full details on MWN are reported in (Pianta et al., 2002).

- the set of direct hypernyms of *s* in the taxonomy hierarchy in MWN;
- the set of synset words in MWN standing in the relation of *nearest synonyms* with *s*;
- the set of synset words in MWN composing the manually disambiguated glosses of *s* from the "Princeton Annotated Gloss Corpus"[5]. To extract the corresponding Italian synset(s), we have ported MWN to WN 3.0;
- the set of synset words in MWN composing the gloss of *s* in Italian (when available);
- for verbs, the set of synset words in MWN standing in the relations of *entailment/is_entailed*, *causes/is_caused* with *s*;
- for nouns, the set of synset words in MWN standing in the relations of *part_of/has_part*, *has_member/is_member* with *s*.

The alignment of senses is based on the notion of lexical overlap. We used `Text::Similarity` v.0.09 module[6], and in particular the method `Text::Similarity::Overlaps`, to obtain the overlap value between two bags of words of s_w. Text similarity is based on counting the number of overlapping tokens between the two strings, normalized by the length of the strings. To overcome the so called "lexical gap" problem (Meyer and Gurevych, 2011), i.e. a reduced number of overlapping words, we have extended the noun sense descriptions of MWN with the Italian Wikipedia glosses extracted from BabelNet (Navigli and Ponzetto, 2012). As for our task, we have retained only those BabelNet entries which have a corresponding synset word in MWN.

3.2. Sense Similarity

In the second approach, Sense Similarity, the basis for sense alignment is the Personalized Page Rank (PPR) algorithm (Eneko and Soroa, 2009) relying on a lexical-semantic knowledge base model as a graph *G* = (V, E) as available in the UKB tool suite[7]. The PPR algorithm ranks the vertices in a graph obtained from a lexical knowledge base according to their importance within the set and assigns stronger initial probabilities to certain kinds of vertices in the graph. The result of the PPR algorithm is a vector. To build the SC Lexicon vectors, we have used two approaches: i.) apply the PPR algorithm on the Italian data using as lexical knowledge base the MWN lexicon; and ii.) apply the PPR algorithm on automatic translations[8] of the SC glosses using as lexical knowledge base WN 3.0. In both cases, the PPR vectors of the SC Lexicon are semantic representations overall the entire MWN or WN synsets of the textual definition of *s*.
As for the MWN synsets, instead of building the PPR vector by means of the lexical items, we have passed to the UKB tool suite the MWN synset id, thus assuming that the MWN synset is already disambiguated. The vector representations of the MWN synsets have been obtained both from the MWN and from the conversion of MWN to WN 3.0.
Finally, given two PPR vectors, namely ppr_{mwn} and ppr_{sc} for the MWN synset w_{syn} and for the SC Lexicon sense w_{sc}, we calculated their cosine similarity. On the basis of the similarity score, the sense pair is considered as aligned or not.

3.3. Results and Evaluation

Each alignment method has been evaluated on its own with respect to Precision, Recall and F-measure over two manually created Gold Standards, one for verbs, which is composed by 44 lemma for a total of 350 aligned sense couples, and one for nouns, which is composed by 46 lemmas for nouns for a total of 166 aligned sense couples. We selected random match (*rand*) as a baseline. The random match works as follows: for the same word *w* in the SC Lexicon and in MWN, it assigns a random SC Lexicon meaning to each synset with *w* as synset word, returning a one-to-one alignment. The identification of the correct aligned pairs has been obtained by applying two types of thresholds with respect to all proposed alignments (the "all_pairs" row in the tables): i.) a simple cut-off at specified values (0.1; 0.2); ii.) the selection of the maximum score (either lesk measure or cosine; row "max_score" in the tables) between each synset *S* and the proposed aligned senses of the SC Lexicon. As for the maximum score threshold, we have retained as good alignments also instances of a tie, thus allowing the possibility of having one MWN synset aligned to more than one SC Lexicon sense.

Lexical Overlap Results Different combinations of the sense representation of a synset have been created. We developed two basic representations: SYN, which is composed by the set of synset words excluding the target word *w* to be aligned, all of its direct hypernyms, the set of synset words in MWN standing in the relation of *nearest synonyms* and the synset words obtained from the "Princeton Annotated Gloss Corpus"; and SREL, which contains all the items of SYN plus the synset words included in the selected set of semantic relations. The results are reported in Table 1 for verbs and Table 2 for nouns.

Lexical Match	P	R	F1
Verb SYN - all_pairs	0.41	0.29	**0.34**
Verb SYN - ≥ 0.1	0.42	0.26	0.32
Verb SYN - ≥ 0.2	0.54	0.11	0.18
Verb SYN - max_score	0.59	0.19	**0.29**
Verb SREL - all_pairs	0.38	0.32	**0.35**
Verb SREL - ≥ 0.1	0.40	0.27	0.32
Verb SREL - ≥ 0.2	0.53	0.11	0.18
Verb SREL - max_score	0.60	0.20	**0.30**
Verb - `rand`	0.15	0.06	0.08

Table 1: Results of Lexical Match for basic sense representation of verbs.

[5]http://wordnet.princeton.edu/glosstag.shtml
[6]http://www.d.umn.edu/ tpederse/text-similarity.html
[7]http://ixa2.si.ehu.es/ukb/
[8]We use Google Translate API.

Lexical Match	P	R	F1
Noun SYN - all_pairs	0.52	0.59	**0.55**
Noun SYN - ≥ 0.1	0.58	0.41	0.48
Noun SYN - ≥ 0.2	0.71	0.16	0.26
Noun SYN - max_score	0.69	0.42	**0.52**
Noun SREL - no_threshold	0.49	0.60	**0.54**
Noun SREL - ≥ 0.1	0.60	0.40	0.48
Noun SREL - ≥ 0.2	0.71	0.13	0.22
Noun SREL - max_score	0.69	0.42	**0.52**
Noun - `rand`	0.17	0.12	0.14

Table 2: Results alignment of Lexical Match for basic sense representation of nouns.

Both basic synset configurations, SYN and SREL, outperform the baseline `rand` for both parts of speech in analysis. The alignment of nouns performs better than that for verbs in both sense representations and with all filtering methods. A manual exploration of the data for verbs and nouns has highlighted that, on the one hand, we suffer from data sparseness on the SC Lexicon side as no extension of the sense description of the glosses is possible, and, on the other hand, that senses are described in ways that are semantically equivalent but with different lexical items. This explains the low Recall figures for both parts-of-speech. The difference in performance of the SREL configuration with respect to the SYN configuration for both parts-of-speech is not statistically significant ($p > 0.05$), suggesting that the impact of additional semantic relations (as encoded in MWN) is limited. Both for verbs and nouns we decided to select the SYN basic configuration as the best sense representation because it has a simpler bag-of-words and better Precision.

To improve the results, we have extended the SYN representation for nouns with the lexical items in the corresponding glosses of BabelNet (+BABEL). The results are illustrated in Table 3.

Lexical Match	P	R	F1
Noun SYN+BABEL - all_pairs	0.47	0.66	**0.56**
Noun SYN+BABEL - ≥ 0.1	0.58	0.40	0.47
Noun SYN+BABEL - ≥ 0.2	0.69	0.12	0.21
Noun SYN+BABEL - max_score	**0.69**	0.44	**0.55**

Table 3: Results for Lexical Match alignment with extensions with BabelNet data.

The extension of the basic sense representations with additional data is positive, namely for Recall at a low or null cost for Precision for all filtering methods. It is interesting to notice that for nouns both the two basic sense descriptions, SYN and SREL, and the SYN+BABEL configuration have comparable F1 values between the no threshold and the maximum score data. Nevertheless, the filtering based on the maximum score improves the quality of the proposed alignment by removing false positives (P=0.69 for SYN, SREL, and SYN+BABEL) without impacting on the number of good instances retrieved (R=0.42 for SYN and SREL, R=0.44 for SYN+BABEL).

Sense Similarity Results The results for the Sense Similarity obtained from the Personalized Page Rank algorithm on the basis of the method described in Section 3.2. are illustrated in Table 4 for the vectors obtained from MWN and in Table 5 for those obtained from WN 3.0 (using automatic translation of the SC glosses).

Similarity Measure	P	R	F1
Verb - all_pairs	0.12	0.69	0.20
Verb - ≥ 0.1	0.33	0.19	**0.24**
Verb - ≥ 0.2	**0.41**	0.13	0.20
Verb - max_score	0.34	0.14	0.20
Verb - `rand`	0.15	0.06	0.08
Noun - all_pairs	0.20	0.64	0.21
Noun - ≥ 0.1	0.42	0.28	**0.33**
Noun - ≥ 0.2	**0.51**	0.18	0.27
Noun - max_score	0.38	0.30	0.34
Noun - `rand`	0.17	0.12	0.14

Table 4: Results for Similarity Score based on MWN.

Similarity Measure	P	R	F1
Verb - all_pairs	0.10	0.9	0.19
Verb - ≥ 0.1	0.47	0.25	**0.32**
Verb - ≥ 0.2	**0.66**	0.16	0.26
Verb - max_score	0.42	0.20	0.27
Verb - `rand`	0.15	0.06	0.08
Noun - all_pairs	0.12	0.94	0.21
Noun - ≥ 0.1	0.52	0.32	**0.40**
Noun - ≥ 0.2	**0.77**	0.21	0.33
Noun - max_score	0.42	0.38	0.40
Noun - `rand`	0.17	0.12	0.14

Table 5: Results for Similarity Score based on WN 3.0.

Similarly to the Lexical Match, the Sense Similarity outperforms the baseline `rand` both when using MWN as lexical knowledge base and when using WN 3.0. Overall, the differences in performance with the Lexical Match results are not immediate. The differences are strictly related to the different nature of the sense descriptions, i.e. a *semantic* representation based on a lexical knowledge graph, which can catch semantically related items out of the scope for the Lexical Match approach. We want to point out that the performances of this approach are strictly dependent on two interrelated aspects: i.) the coverage of the dictionary entries used by the lexical knowledge base, and ii.) the set of relations which are represented in the lexical knowledge base.

Concerning the use of MWN as lexical knowledge base, the overall results are lower than those obtained for Lexical Overlap, and the use of automatic translations and WN 3.0. Comparable results with Lexical Overlap are obtained only for Recall with no filtering (`all_pairs` row in Table 4) both for nouns and verbs. As a manual error analysis has shown, these results are stricly related to the structure of MWN as lexical knowlege base, i.e. poor coverage in terms of relations and entries in the dictionary. For instance, we identified that most of the vectors for SC verbs have not been created due to the lack of entries in the lexical knowlege base dictionary. This aspect also support our previous observations on the results of Lexical Match.

On the contrary, the use of WN 3.0 provides interesting figures. Although not all MWN synsets have a corresponding entry in WN 3.0, the size of the English dictionary and the relations among the entries in the WN 3.0 graph provide more aligned pairs with respect to MWN, as shown by the figures for Recall in Table 5. In particular, by observing the Recall values for no threshold filtering (row `all_pairs` Table 5), almost all aligned sense pairs of the gold are retrieved, outperforming both the Lexical Match and the similarity with MWN. As for the filtering methods, figures for verbs and nouns show that the simple cut-off thresholds provide better results with respect to the maximum score. Such a better performance of the simple cut-off thresholds with respect to the maximum score is due to the fact that aligning senses by means of semantic similarity provides a larger set of alignments and facilitates the identification of multiple alignments, i.e. one-to-many.

As for verbs, the best the best F1 score (F1=0.32) is obtained when setting the cosine similarity to 0.1, though Precision is less than 0.50. When compared with threshold value of 0.1 of the Lexical Match and similarity with MWN, the similarity with WN 3.0 yields the best Precision (P=0.47 *vs*. P=0.42 for Verb SYN, P=0.40 for Verb SREL and P=0.33 for similarity with MWN). Similar observations can be done when the threshold is set to 0.2. In this latter case, similarity with WN 3.0 yields the best Precision score with respect to all other filtering methods and the Lexical Match results obtained with maximum score (P=0.66 *vs*. P=0.59 for Verb SYN, P=0.60 for Verb SREL and P=0.41 for similarity with MWN).

The results for nouns are different though in line with those for verbs. Apparently, the similarity with WN 3.0 has better results for F1 only with respect to similarity with MWN and lower values with respect to all Lexical Match sense configurations and filtering methods, including the no threshold score of the basic sense descriptions (respectively, F1=0.55 for SYN, F1=0.54 for SREL, F1=0.21 for similarity with WN 3.0). However, when maximizing the Precision for the similarity with WN 3.0 (threshold 0.2), the algorithm provides better performances (F1=0.33) with respect to Lexical Match on the same filtering method, minimizing the drop of Recall (R=0.21; +0.09 with respect to SYN+BABEL with same threshold; + 0.08 with respect to SREL; +0.05 with respect to SYN, respectively).

Merging Lexical Match and Similarity with WN 3.0

The methods used for aligning senses in the two lexico-semantic resources differ in nature both with respect to the creation of the sense descriptions (simple bag of words *vs*. semantic representation) and to the ways with which the alignment pairs are extracted and computed. As a strategy to improve the results, we conducted a further alignment experiment by merging together the results obtained from the best sense descriptions and best filtering methods for Lexical Match and Sense Similarity, namely similarity with WN 3.0. We considered Precision and F1 scores to identify the best results. This led us to select the i.) the SYN sense description filtered with maximum score for verbs (P=0.59, F1=0.29); ii.) the SYN+BABEL sense description filtered with maximum score for nouns; iii.) the similarity with WN 3.0 with the cut-off threshold at 0.2 both for verbs and nouns. The results of the merging are illustrated in Table 6.

Merged	**P**	**R**	**F1**
Verb - SYN+SimWN30_02	0.61	0.38	**0.47**
Noun - SYN+BABEL+SimWN30_02	0.67	0.61	**0.64**

Table 6: Results for automatic alignment merging the best results from Lexical Match and Sense Similarity.

The merging has a positive impact on the alignments of both parts-of-speech, signaling that different methods are focused on capturing different portions of the data. Global F1 scores are improved both for nouns and verbs. Nevertheless, the figures for Precision are still not totally satisfactory. In both cases the performance gains originate from the higher precision of the similarity approach with WN 3.0 with automatically translated glosses which minimizes the limits of the Italian section of MWN and of the SC Lexicon.

4. Conceptual Interoperability with Verb Shallow Frames

In order to enhance conceptual interoperability on other levels of linguistic analysis for Senso Comune, we have automatically extracted verbal shallow frames (VSFs) from a sense annotated corpus in Italian, namely the MultiSemCor Corpus v1.0, a parallel corpus of English and Italian annotated with WN senses. The final goal is, for each sense annotated verb in the Italian section of MultiSemCor, to extract all available corpus-based example and automatically acquire VSFs. This operation will allow us: i.) to provide a starting set of verb structures associated with (M)WN senses and corpus-based examples for the development of layered annotations in the line of VerbNet; ii.) to investigate on the correlation between verb senses and different VSFs; and, finally, iii.) to experiment on the improvement of verb sense alignment. In this paper we will focus on the description and a preliminary evaluation of the first aspect, that is on the acquisition of VSF structures from the MultiSemCor Corpus[9].

We assumed as a VSF structure the syntactic complements of the verb, with no distinction between arguments and adjuncts, and the semantic type of the complement filler(s). An example of an SFS is reported in Example 1.

1. *Marco ha comprato un libro*.
   [Marco bought a book.]
   Verb: *comprare* [to buy]
   SFS: SUBJ[person] OBJ[artifact]

The original gold standard used for evaluating the WSA task has been extended to include 52 seed lemmas selected according to frequency and patterns in terms of semantic and syntactic features[10]. For each seed verb, we have extracted all its corresponding synsets in MWN and synset words. This has lead us to identify a total of 167 unique verb lemmas and a total of 417 different synsets. We

[9]Experiments on the use of VSFs for improving the alignment of verb senses are described in (Caselli et al., 2013).

[10]A subset of these verbs have been taken from (Jezek and Quochi, 2010)

then extracted from MultiSemCor all associated sentences which contained an annotated instance of the 417 synsets for a total of 4,820 instances.
The extraction of the VSFs has been obtained as follows:

- MultiSemCor sentences have been parsed with a state-of-the-art dependency parser (Attardi and Dell'Orletta, 2009);
- for each verb lemma, we have automatically extracted all its syntactic complements standing in a dependency relation of argument or complement, together with the lemma of the slot filler;
- nominal lemmas of syntactic complements have been automatically assigned with one of the 26 semantic types composing the WN supersenses (i.e. *noun.artifact*; *noun.object* etc. (Ciaramita and Johnson, 2003)) on the line of (Lenci et al., 2012). For each nominal filler, we selected the most frequent WN supersense. Sense frequency had been computed on the basis of MultiSemCor. In case a polysemous noun lemma was not present in the MultiSemCor data or its senses have the same frequency, all associated WN supersenses were assigned. As for verbal fillers, we assigned the generic semantic type of "*verb.eventuality*". Finally, in case a lemma filler of a syntactic complement is not attested in MWN such as a pronoun or a missing synset word, no values is assigned and the VFS is not constructed. Optionally, when the noun filler was annotated with a synset in MultiSemCor, we have associated it to its corresponding WN supersense.

The information thus collected has been stored in a theory- and model neutral format, as illustrated in Example 1, which is compatible with other representation formats for VSF structure such as that in the PAROLE/SIMPLE/CLIPS lexicon (Ruimy et al., 2003) and with current research activities in Senso Comune on the annotation of SC Lexicon examples of usage for verbs (Chiari et al., 2013).
As for the extraction of the VSF structures, we have used a modified version of the system described in (Caselli et al., 2012), which reported an overall F-measure on the acquisition of VSFs from corpus data of 0.601 and a Precision of 0.65. The original system has a filtering mechanism based on maximum likelihood estimate (MLE) and percentage on verb frequency (PVF) which is used to exclude incorrect VSFs on the basis of the frequency of the VFS and the verb lemma in the corpusused for the acquisition. In our version, we have excluded this filtering mechanism, as we are working on verb senses and not lemmas. As a matter of fact, each verb sense has a limited number of annotated instances. For example, the verb *aprire* [to open] in MultiSemCor has been annotated in 42 sentences with 11 different synsets. By excluding the filtering mechanism, we consider as valid all extracted VSFs. Nevertheless, it can be the case that the extracted VSFs is not correct. In order to provide a preliminary evaluation of the quality of the extracted VSFs with this approach, we have adopted the following method: we have extracted with the original version of the system in (Caselli et al., 2012), all VSFs for the 167 lemmas from a parsed version of the La Repubblica Corpus (Baroni et al., 2004). This has provided us with a repository of VSFs associated to verb lemmas. We then computed the accuracy of the extracted VSFs from the MultiSemCor sentences with those contained in the repository from La Repubblica. The accuracy provides us with a percentage measure of the correctly identified VSFs attested in the repository and of the non-attested ones.
From the 4,820 sentences of the MultiSemCor corpus in analysis, we extracted 3,295 VSFs tokens for the couples verb lemma-annotated sense. The 3,295 VSFs tokens correspond to 418 VSFs types. On top of this 3,295 acquired VSF tokens we computed their accuracy with respect to the La Repubblica repository as described above. The results are illustrated in Table 7.

VSF Tokens	Attested	Not Attested
3,295	2,232 (68%)	1,063 (32%)

Table 7: Accuracy on the extracted VSFs with respect to the La Repubblica repository

On a manual exploration of the non attested VSFs, we observe that most of them are due to parsing errors and VSFs which have been excluded by the filtering mechanism of the system. We are currently investigating on the application of crowdsourcing techniques on the Senso Comune platform to perfoem a post-processing analysis on the non-attested VFSs.

5. Conclusions and Future Work

This paper focuses on methods for automatically enrich the Senso Comune platform in the perspective of creating a robust and interoperable lexico-semantic resource for Italian. Two tasks have been tackled: i.) aligning senses between MWN and the SC Lexicon, and ii.) automatically acquire VSFs from a sense annotated corpus.
As for the sense alignment task, the lack of Italian glosses in MWN and the absence of any kind of structured information in the SC Lexicon dictionary are two major challenges for the applications of state-of-the-art techniques for sense alignment. Two different approaches have been experimented: Lexical Match and Sense Similarity obtained from Personalized Page Rank. In all cases, when filtering the data we are facing low scores for Recall which point out issues namely related to data sparseness in our lexica. When comparing the results of the two approaches, we can observe that: i.) the Sense Similarity by means of automatic gloss translations plus WN 3.0 as lexical knowledge base yields the best Precision both with respect to Lexical Match and to Sense Similarity with MWN; ii.) Lexical Match, with a simple sense description configuration (i.e. the SYN configurations for verbs and nouns), is still a powerful approach; the exploitation of additional semantically related items (e.g. SREL for verbs) or additional sense descriptors (e.g. SYN+BABEL for nouns), though good in principle, has a limited contribution to solve the lexical gap problem in our case and highlights differences in the way word senses are encoded in the two lexica; iii.) Sense Similarity with automatic gloss translations and WN 3.0 performs better than Sense Similarity with MWN, pointing out

that MWN as a lexical knowledge base has a lower coverage and that seems worst that having not perfect translations; and iv.) Sense Similarity and Lexical Match appears to qualify as complementary methods for achieving reliable sense alignments. Provided the limits of the two lexica, we have obtained satisfying results both for verb (F1=0.47) and noun sense alignment (F1=0.64). Nevertheless, we consider that there is still room for improving the results, namely in terms of Precision.

The acquisition of VSFs is another important task for achieving conceptual interoperability. The method used for the acquisition of the VSFs is reliable (accuracy=68%), though changes in the acquisition system should be tackled to deal with missing VSFs. The limited amount of not attested VSFs (32%) reduce the manual effort in the post-processing phase and will allow to experiment with crowdsourcing methods for the annotation of VSFs. Furthermore, the data will provide a basis for the development of a VerbNet-like resource for Italian.

Future work will concentrate on two different aspects. We are currently investigating methods for automatically assign WN Domains to the SC Lexicon entries. Preliminary results are encouraging. The availability of WN Domain can be used to filter the proposed alignments and remove most cases of false positive data, thus increasing the Precision. Furthermore, we aim at importing the ontological classes of SC in MWN for bootstrapping better sense descriptions and investigating additional taxonomical errors in the (M)WN hierarchy with respect to those identified in (Alvez et al., 2008). As for the VSFs, on the one hand, we are planning to integrate the automatically acquired data from MultiSemCor with the manual annotation of the SC example of usage for the verbal entries (Chiari et al., 2013), and, on the other hand, we will experiment on the development of methods for the "leaking" of the Semantic Roles associated to WN senses in VerbNet to the corresponding Italian verb senses and syntactic structures.

6. References

Alvez, J., Atserias, J., Carrera, J., Climent, S., Laparra, E., Oliver, A., and Rigau, G. (2008). Complete and consistent annotation of wordnet using the top concept ontology. In *Proceedings of the Sixth International conference on Language Resources and Evaluation (LREC-08)*.

Attardi, G. and Dell'Orletta, F. (2009). Reverse revision and linear tree combination for dependency parsing. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, Boulder, Colorado.

Baroni, M., Bernardini, S., Comastri, F., Piccioni, L., Volpi, A., Aston, G., and Mazzoleni, M. (2004). Introducing the "`la Repubblica`" corpus: A large, annotated, `TEI(XML)`-compliant corpus of newspaper italian. In *Proceedings of the Fourth International conference on Language Resources and Evaluation (LREC-04)*.

Bentivogli, L. and Pianta, E. (2005). Exploiting parallel texts in the creation of multilingual semantically annotated resources: the MultiSemCor Corpus. *Natural Language Engineering*, 11:247–261, 8.

Buitelaar, P., Cimiano, P., Haase, P., and Sintek, M. (2009). Towards linguistically grounded ontologies. In *The semantic web: research and applications*, pages 111–125. Springer.

Caselli, T., Rubino, F., Frontini, F., Russo, I., and Quochi, V. (2012). Customizable scf acquisition in italian. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, Istanbul, Turkey.

Caselli, T., Vieu, L., Carlo, S., and Vetere, G. (2013). Aligning verb senses in two italian lexical semantic resources. In *Joint Symposium on Semantic Processing.*, page 33.

Chiarcos, C., McCrae, J., Cimiano, P., and Fellbaum, C. (2013). Towards open data for linguistics: Linguistic linked data. In *New Trends of Research in Ontologies and Lexical Resources*, pages 7–25. Springer.

Chiari, I., Gangemi, A., Jezek, E., Oltramari, A., Vetere, G., and Vieu, L. (2013). An open knowledge base for italian language in a collaborative perspective. In *Proceedings of DH-case13, Collaborative Annotations in Shared Environments: metadata, vocabularies and techniques in the Digital Humanities*.

Ciaramita, M. and Johnson, M. (2003). Supersense tagging of unknown nouns in WordNet. In Collins, M. and Steedman, M., editors, *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, pages 168–175.

De Mauro, T. (2000). *Grande dizionario italiano dell'uso*. Utet.

Eneko, A. and Soroa, A. (2009). Personalizing PageRank for Word Sense Disambiguation. In *Proceedings of the 12th conference of the European chapter of the Association for Computational Linguistics (EACL-2009)*, Athens, Greece.

Fang, A. C. (2012). Creating an interoperable language resource for interoperable linguistic studies. *Language resources and evaluation*, 46(2):327–340.

Jezek, E. and Quochi, V. (2010). Capturing coercions in texts: a first annotation exercise. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, pages 1464–1471, Valletta, Malta. European Language Resources Association (ELRA).

Lenci, A., Lapesa, G., and Bonansinga, G. (2012). Lexit: A computational resource on italian argument structure. In *Proceedings of LREC, 8th International Conference on Language Resources and Evaluation*, Istanbul, Turkey.

Masolo, C., Gangemi, A., Guarino, N., Oltramari, A., and Schneider, L. (2002). Wonderweb deliverable D17: the wonderweb library of foundational ontologies. Technical report.

Matuschek, M. and Gurevych, I. (2013). Dijkstra-wsa: A graph-based approach to word sense alignment. *Transactions of the Association for Computational Linguistics (TACL)*, 2:to appear.

Meyer, M. and Gurevych, I. (2011). What psycholinguists know about chemistry: Aligning Wiktionary and WordNet for increased domain coverage. In *Proceedings of the 5th International Joint Conference on Natural Language Processing (IJCNLP)*.

Navigli, R. and Ponzetto, S. P. (2012). BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

Navigli, R. (2006). Meaningful clustering of senses helps boost word sense disambiguation performance. In *Proceedings of the* 44^{th} *Annual Meeting of the Association for Computational Linguistics joint with the* 21^{st} *International Conference on Computational Linguistics (COLING-ACL)*, Sydney, Australia.

Niemann, E. and Gurevych, I. (2011). The peoples web meets linguistic knowledge: Automatic sense alignment of Wikipedia and WordNet. In *Proceedings of the 9th International Conference on Computational Semantics*, pages 205–214, Singapore, January.

Oltramari, A., Vetere, G., Chiari, I., Jezek, E., Zanzotto, F. M., Nissim, M., and Gangemi, A. (2013). Senso Comune: A collaborative knowledge resource for italian. In Gurevych, I. and Kim, J., editors, *The Peoples Web Meets NLP, Theory and Applications of Natural Language Processing*, pages 45–67. Springer-Verlag, Berlin Heidelberg.

Pianta, E., Bentivogli, L., and Girardi, C. (2002). MultiWordNet: developing an aligned multilingual database. In *First International Conference on Global WordNet*, Mysore, India.

Pianta, E., Girardi, C., and Zanoli, R. (2008). TextPro Tool Suite. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC-08)*, volume CD-ROM, Marrakech, Morocco. European Language Resources Association (ELRA).

Ruimy, N., Monachini, M., Gola, E., Calzolari, N., Fiorentino, M. C. D., Ulivieri, M., and Rossi, S. (2003). A computational semantic lexicon of italian: *SIMPLE*. *Linguistica Computazionale XVIII-XIX, Pisa*, pages 821–64.

Shi, L. and Mihalcea, R. (2005). Putting pieces together: Combining framenet, verbnet and wordnet for robust semantic parsing. In *Computational Linguistics and Intelligent Text Processing*, pages 100–111. Springer.

Vetere, G., Oltramari, A., Chiari, I., Jezek, E., Vieu, L., and Zanzotto, F. M. (2012). Senso Comune, an open knowledge base for italian. *TAL-Traitement Automatique des Langues, Special Issue*, 52(3):217–243.

Witt, A., Heid, U., Sasaki, F., and Sérasset, G. (2009). Multilingual language resources and interoperability. *Language Resources and Evaluation*, 43(1):1–14.