10.12871/CLICIT2014119

Senso Comune as a Knowledge Base of Italian language: the Resource and its Development

Tommaso Caselli
VU Amsterdam
t.caselli@vu.nl

Isabella Chiari
Università di Roma 'Sapienza'
isabella.chiari@uniroma1.it

Aldo Gangemi
CNR ISTC
agangemi@loa-cnr.it

Elisabetta Jezek
Università di Pavia
jezek@unipv.it

Alessandro Oltramari
Carnegie Mellon University
aoltrama@andrew.cmu.edu

Guido Vetere
IBM Italia
gvetere@it.ibm.com

Laure Vieu
CNRS IRIT
vieu@irit.fr

Fabio Massimo Zanzotto
Università di Roma 'Tor Vergata'
zanzotto@info.uniroma2.it

Abstract

English. *Senso Comune* is a linguistic knowledge base for the Italian Language, which accommodates the content of a legacy dictionary in a rich formal model. The model is implemented in a platform which allows a community of contributors to enrich the resource. We provide here an overview of the main project features, including the lexical-ontology model, the process of sense classification, and the annotation of meaning definitions (glosses) and lexicographic examples. Also, we will illustrate the latest work of alignment with MultiWordNet, to illustrate the methodologies that have been experimented with, to share some preliminary result, and to highlight some remarkable findings about the semantic coverage of the two resources.

Italiano. *Senso Comune è una base di conoscenza della lingua italiana, che offre il contenuto di un dizionario tradizionale in un ricco modello formale. Il modello è implementato in una piattaforma che consente di arricchire la risorsa ad una comunità di contributori. Qui forniamo una panoramica delle principali caratteristiche del progetto, compreso il modello lessicale-ontologico, il processo di classificazione dei sensi, l'annotazione delle definizioni (glosse) ed degli esempi d'uso lessicografici. Tratteremo inoltre del lavoro di allineamento con MultiWordNet, illustrando le metodologie che sono state sperimentate, e riportando alcune considerazioni circa la copertura semantica delle due risorse.*

1 Introduction

Senso Comune[1] is an open, machine-readable knowledge base of the Italian language. The lexical content has been extracted from a monolingual Italian dictionary[2], and is continuously enriched through a collaborative online platform. The knowledge base is freely distributed. *Senso Comune* linguistic knowledge consists in a structured lexicographic model, where senses can be qualified with respect to a small set of ontological categories. *Senso Comune*'s senses can be further enriched in many ways and mapped to other dictionaries, such as the Italian version of MultiWordnet, thus qualifying as a linguistic Linked Open Data resource.

1.1 General principles

The *Senso Comune* initiative embraces a number of basic principles. First of all, in the era of user generated content, lexicography should be able to build on the direct witness of native speakers. Thus, the project views at linguistic knowledge acquisition in a way that goes beyond the exploitation of textual sources. Another important assumption is about the relationship between language and ontology (sec. 2.1). The correspondence between linguistic meanings, as they are listed in dictionaries, and ontological categories, is not direct (if any), but rather *tangential*. Linguistic senses commit to the existence of various

[1] www.sensocomune.it

[2] T. De Mauro, Grande dizionario italiano dell'uso (GRADIT), UTET 2000

kinds of entities, but should not be in general confused with (and collapsed to) logical predicates directly interpretable on these entities. Finally, we believe that, like the language itself, linguistic knowledge should be owned by the entire community of speakers, thus they are committed to keep the resource open and fully available.

2 Senso Comune Essentials

2.1 Lexicon and ontology

In compliance with recent trends of research in integrating ontologies and lexical resources (see e.g. (Oltramari et al., 2013) and (Prévot et al., 2010)) *Senso Comune* model includes a lexicon and an ontology as independent semantic layers. Instead of providing *synsets* with formal specifications aimed at qualifying them as ontological classes (Gangemi et al., 2003), *Senso Comune* adopts a notion of *ontological commitment*, which can be summarized as follows:

If the sense S commits to ($\mapsto$) the concept C, then there are entities of type C to which occurrences of S may refer to.

$$(S \mapsto C) \Leftrightarrow \exists s, c | S(s) \wedge C(c) \wedge refers\_to(s, c)$$

This way, linguistic senses are not modelled as logical predicates to be directly interpreted with respect to individuals in some domain of quantification, but rather as *semiotic objects* that occur in texts or communication acts, whose relationship with other real world entities is mediated by cognitive structures, emotional polarity and social interactions.

As a consequence of this model, lexical relations such as synonymy, which hold among senses, do not bear any direct ontological import; conversely, ontological axioms, such as disjointness, do not have immediate linguistic side-effects. This approach allows senses of different types to be freely put into lexical relations, without the need of assigning the same (complex) type to every member of the synonymy relation; on the other hand, it prevents the system from directly inferring ontological relations out of linguistic evidences, which might be a limitation in many cases. Anyway, if the equivalence of linguistic senses to logic predicates is desired (e.g. for technical, monosemic portions of the dictionary), this condition can be specifically formalized and managed.

2.2 Sense classification

Meanings from De Mauro's core Italian lexicon have been clustered and classified according to ontological categories belonging to *Senso Comune* model, through a supervised process we called TMEO, a tutoring methodology to support sense classification by means of interactive enrichment of ontologies (Oltramari, 2012). TMEO is based on broad foundational distinctions derived from a simplified version of DOLCE[3] (Masolo et al., 2002) (Chiari et al., 2013). The overarching goal is to support users that, by design, have only access to the lexical level of the resource, in the task of selecting the most adequate category of the *Senso Comune* ontology as the super-class of a given lexicalized concept: different answer paths lead to different mappings between the lexical and the ontological layer of *Senso Comune* knowledge base.

Ongoing work on TMEO focuses on extending the coverage of the methodology and refining both the category distinctions in the ontology and the questions in the decision tree. In a previous experiment reported in (Chiari et al., 2010), we observed that users have a high degree of confidence and precision in classifying the concepts referring to the physical realm, while they face several problems in identifying abstract notions like 'company', 'text', 'beauty', 'duration', 'idea', etc. Accordingly, the new scheme, already tested in our last experiment (Jezek et al., 2014) summarized below, mainly improves the *Senso Comune* ontology in the abstract realm. It substitutes the too vague category `Idea` with the more generic `SocialOrMentalObject`, within which `InformationObject` and `Organization` are distinguished subcategories. In addition, the remaining abstract categories `TemporalQuality`, `Quality` and `Function` are complemented and grouped under a more general category `PropertyOrRelation`. Finally, we added the possibility to distinguish, for each category, a singular and a collective sense, thus allowing to annotate the main senses of the lemmas 'popolo' (people) and 'gregge' (herd) with the categories `Person` and `Animal` (adding a 'collective' tag). The results are a richer taxonomy and better organized decision tree.

[3] `http://www.loa.istc.cnr.it/old/DOLCE.html`

2.3 Annotation of lexicographic examples and definitions

Ongoing work in *Senso Comune* focuses on manual annotation of the usage examples associated with the sense definitions of the most common verbs in the resource, with the goal of providing *Senso Comune* with corpus-derived verbal frames. The annotation task, which is performed through a Web-based tool, is organized in two main subtasks. The first (task 1) consists in identifying the constituents that hold a relation with the target verb in the example and to annotate them with information about the type of phrase and grammatical relation. In semantic annotation (task 2), users are asked to attach a semantic role, an ontological category and the sense definition associated with the argument filler of each frame participant in the instances. For this aim, we provide them with a hierarchical taxonomy of 24 coarse-grained semantic roles based on (Bonial et al., 2011), together with definitions and examples for each role, as well as decision trees for the roles with rather subtler differences. The TMEO methodology is used to help them selecting the ontological category in a new simplified ontology based on *Senso Comune*'s top-level. For noun sense tagging, the annotator exploits the senses already available in the resource. Drawing on the results of the previous experiment on nouns senses, we allow multiple classification in all the three semantic subtasks, that is, we allow the users to annotate more than one semantic role, ontological category and sense definition for each frame participant. Up to now we performed two pilot experiments to release the beta version of the annotation scheme. The results of IA agreement are very good for the syntactic dependency annotation task and fair for the semantic task, the latter especially so since these tasks are notoriously difficult (see (Jezek et al., 2014) for details). Once completed, the annotated data will be used to conduct an extensive study of the interplay between thematic role information and ontological constraints associated with the participants in a frame; to refine the ontologisation of nouns senses in *Senso Comune* by assigning ontological classes to nouns in predicative context instead of nouns in isolation; to investigate systematic polysemy effects in nominal semantics on a quantitative basis. Our long-term goal is to enrich the resource with a rich ontology for verb types, informed by the empirical data provided by the annotated corpus.

3 Word Sense Alignment: Towards Semantic Interoperability

As a strategy to enrich the Senso Comune Lexicon (SCL) and make it interoperable with other Lexico-semantic resources (LSRs), two experiments of Word Sense Alignment (WSA) have been conducted: a manual alignment and an automatic one. WSA aims at creating a list of pairs of senses from two (or more) lexical-semantic resources where each pair of aligned senses denotes the same meaning (Matuschek and Gurevych, 2013). The target resource for the alignment is MultiWordNet (MWN) (Pianta et al., 2002).
SCL and MWN are based on different models[4]. The alignment aims at finding a semantic portion common to the set of senses represented in SCL by the conjunction of glosses and usage examples and in MWN by the synset words and their semantic relationships (hypernyms, hyponyms, etc.). Since semantic representation in the form of lexicographic glosses and in the form of synsets cannot be considered in any respect homomorphic the procedure of alignment is not biunique in any of the two directions. Thus, there are single SCL glosses aligned to more than one MWN synsets and single MWN synsets aligned with more than one SCL gloss. Another goal of the alignment experiments is the integration of high quality Italian glosses in MWN, so as to make available an enhanced version of MWN to NLP community, which could help improving Word Sense Disambiguation (WSD) and other tasks.

3.1 Manual aligment

On going work on the manual alignment of SCL and MWN synsets aims at providing associations between SCL glosses and synsets for all 1,233 nouns labelled as belonging to the basic vocabulary. The alignment is performed through the online platform that allows for each SCL word sense the association with one or more MWN synset.
At the time of this writing, 584 lemmas of SCL have been processed for manual alignment, for a total of 6,730 word senses (glosses), about 3.64 average word senses for each lemma. The alignment involves all SCL word senses, including

[4]Readers are referred to (Vetere et al., 2011) and (Caselli et al., 2014) for details on the two resources and their differences.

word senses not labelled as fundamental (about 29% of all word senses). Preliminary results show that only 2,131 glosses could be aligned with at least one MWN synset (31.7%) and 2,187 synsets could be aligned to at least one gloss. Exclusively biunique relationships among SCL glosses and MWN synsets involve 1,093 glosses. Each SCL gloss is associated to one synset in 1,622 cases (76.1%), to two synsets in 367 cases (17.2%), to three synsets 108 cases (5%), to four 25 (1,1%), to five in four cases, to six in three cases and to seven synsets in one case. While on the other side each MWN synset is associated to one SCL gloss in 1,681 cases (76.8%), to two glosses in 400 cases (18.2%), to three glosses in 85 cases (3.8%), to four in 17 cases, to five in three cases, and to six glosses in one case. The picture portrayed by the asymmetry of relationship between the granularity of SCL and MWN appears very similar, meaning that there is no systematic difference in the level of detail in the two resources aligned, as far as this preliminary analysis reveals. Attention should be drawn to the fact that biunique associations do not directly entail that the semantic representation deriving from the SCL gloss and the MWN synset are semantically equivalent or that they regard the same set of senses. These association only indicate that there is no other gloss or synset that can properly fit another association procedure. Levels of abstraction can be significantly different. Furthermore, as data show, there is a large number of SCL glosses not aligned to any MWN synset, and vice versa. This mismatch probably derives from the fact that MWN synsets are modelled on the English WN. Many WN synsets could be aligned to Italian senses outside the basic vocabulary; however, in general, we think that this mismatch simply reflects the semantic peculiarity of the two languages.

3.2 Automatic alignment

We conducted two automatic alignment experiments by applying state-of-the-art WSA techniques. The first technique, Lexical Match, aims at aligning the senses by counting the number of overlapping tokens between two sense descriptions, normalized by the length of the strings. We used `Text::Similarity` v.0.09 The second technique, Sense Similarity, is based on computing the cosine score between the vector representations of the sense descriptions. Vector representations have been obtained by means of the Personalized Page Rank (PPR) algorithm (Agirre et al., 2014) with WN30 extended with the "Princeton Annotated Gloss Corpus" as knowledge base[5]. The evaluation of the automatic alignments is performed with respect to two manually created Gold Standards, one for verbs and one for nouns, by means of standard Precision (P), Recall (R) and F1 score. The verb Gold Standard contains 350 sense pairs over 44 lemmas, while the noun Gold Standard has 166 sense pairs for 46 lemmas. The two gold standards have been independently created with respect to the manual alignment described in Section 3.1 and took into account only fundamental senses. Concerning the coverage of in terms of aligned entries, as for verbs MWN covers 49.76% of the SCDM senses while for nouns MWN covers 62.03% of the SCDM senses. The best results in terms of F1 score have been obtained by merging the outputs of the two approaches together, namely we obtained an F1 equals to 0.47 for verbs (P=0.61, R=0.38) and of 0.64 for nouns (P=0.67, R=0.61).

4 Conclusion

In this paper, we have introduced *Senso Comune* as an open cooperative knowledge base of Italian language, and discussed the issue of its alignment with other linguistic resources, such as WordNet. Experiments of automatic and manual alignment with the Italian MultiWordNet have shown that the gap between a native Italian dictionary and a WordNet-based linguistic resource may be relevant, both in terms of coverage and granularity. While this finding is in line with classic semiology (e.g. De Saussure's *principle of arbitrariness*), it suggests that more attention should be paid to the semantic peculiarity of each language, i.e. the specific way each language constructs a conceptual view of the World. One of the major features of *Senso Comune* is the way linguistic senses and ontological concepts are put into relation. Instead of equalising senses to concepts, a formal relation of *ontological commitment* is adopted, which weakens the ontological import of the lexicon. Part of our future research will be dedicated to leverage on this as an enabling feature for the integration of different lexical resources, both across and within national languages.

[5]Readers are referred to (Caselli et al., 2014) for details on the two methods used and result filtering.

References

Eneko Agirre, Oier Lopez de Lacalle, and Aitor Soroa. 2014. Random walks for knowledge-based word sense disambiguation. *Computational Linguistics*, 40(1):57–84.

C. Bonial, W. Corvey, M. Palmer, V.V. Petukhova, and H. Bunt. 2011. A hierarchical unification of lirics and verbnet semantic roles. In *Semantic Computing (ICSC), 2011 Fifth IEEE International Conference on*, pages 483–489. IEEE.

Tommaso Caselli, Carlo Strapparava, Laure Vieu, and Guido Vetere. 2014. Aligning an italianwordnet with a lexicographic dictionary: Coping with limited data. In *Proceedings of the Seventh Global WordNet Conference*, pages 290–298.

Isabella Chiari, Alessandro Oltramari, and Guido Vetere. 2010. Di che cosa parliamo quando parliamo fondamentale? In S. Ferreri, editor, *Atti del Convegno della Societ di Linguistica Italiana*, pages 185–202, Roma. Bulzoni.

Isabella Chiari, Aldo Gangemi, Elisabetta Jezek, Alessandro Oltramari, Guido Vetere, and Laure Vieu. 2013. An open knowledge base for italian language in a collaborative perspective. In *Proceedings of the 1st International Workshop on Collaborative Annotations in Shared Environment: Metadata, Vocabularies and Techniques in the Digital Humanities*, DH-CASE '13, pages 14:1–14:6, New York, NY, USA. ACM.

Aldo Gangemi, Roberto Navigli, and Paola Velardi. 2003. The ontowordnet project: extension and axiomatization of conceptual relations in wordnet. In *in WordNet, Meersman*, pages 3–7. Springer.

E. Jezek, L. Vieu, F. M. Zanzotto, G. Vetere, A. Oltramari, A. Gangemi, and R. Varvara. 2014. Enriching senso comune with semantic role sets. In *Proceedings of the Tenth Joint ACL-ISO Workshop on Interoperable Semantic Annotation, Reykjavik, Iceland (May 26, 2014)*, pages 88–94.

C. Masolo, A. Gangemi, N. Guarino, A. Oltramari, and L. Schneider. 2002. WonderWeb Deliverable D17: The WonderWeb Library of Foundational Ontologies. Technical report.

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

A. Oltramari, P. Vossen, L. Qin, and E. Hovy, editors. 2013. *New Trends of Research in Ontologies and Lexical Resources*, volume XV of *Theory and Applications of Natural Language Processing*. Springer, Heidelberg.

Alessandro Oltramari. 2012. An introduction to hybrid semantics: The role of cognition in semantic resources. In Alexander Mehler, Kai-Uwe Khnberger, Henning Lobin, Harald Lngen, Angelika Storrer, and Andreas Witt, editors, *Modeling, Learning, and Processing of Text Technological Data Structures*, volume 370 of *Studies in Computational Intelligence*, pages 97–109. Springer Berlin Heidelberg.

Emanuele Pianta, Luisa Bentivogli, and Cristian Girardi. 2002. MultiWordNet: developing an aligned multilingual database. In *First International Conference on Global WordNet*, Mysore, India.

Laurent Prévot, Chu-Ren Huang, Nicoletta Calzolari, Aldo Gangemi, Alessandro Lenci, and Alessandro Oltramari, editors. 2010. *Ontology and the Lexicon*. Cambridge University Press.

Guido Vetere, Alessandro Oltramari, Isabella Chiari, Elisabetta Jezek, Laure Vieu, and Fabio Massimo Zanzotto. 2011. Senso Comune, an open knowledge base for italian. *Traitement Automatique des Langues*, 53(3):217–243.