Taming Sense Sparsity: a Common-Sense Approach

Antonio Lieto, Enrico Mensa and Daniele P. Radicioni
Dipartimento di Informatica
Università degli Studi di Torino
Corso Svizzera 185, 10149 – Torino ITALY
{lieto,mensa,radicion}@di.unito.it

Abstract

We present a novel algorithm and a linguistic resource named CLOSEST after 'Common SEnse STrainer'. The resource contains a list of the main senses associated to a given term, and it was obtained by applying a simple set of pruning heuristics to the senses provided in the NASARI vectors for the set of $15K$ most frequent English terms. The preliminary experimentation provided encouraging results.

Italiano. *In questo lavoro presentiamo un algoritmo e una risorsa linguistica,* ClOSeSt*, che contiene i sensi più rilevanti per i* $15K$ *termini più frequenti del dizionario inglese. L'algoritmo implementato utilizza una risorsa esistente che codifica conoscenza di tipo enciclopedico, e poggia sulla nozione di senso comune per filtrare i possibili sensi associati a ciascun termine. La valutazione preliminare ha fornito risultati incoraggianti in merito alla qualità dei sensi estratti.*

1 Introduction

Many NLP tasks involve word sense disambiguation (WSD) and word sense induction (WSI), and require using lexical resources such as WordNet (Miller, 1995) and BabelNet (Navigli and Ponzetto, 2010) that provide a rich mapping of terms (or word *forms*) onto the corresponding senses (word *meanings*). These widely used resources provide in fact subtle distinctions between the possible senses of a term. It is largely acknowledged that while fine-grained sense distinctions are necessary for some precise tasks (such as machine translation), for other sorts of applications (such as text categorization and information extraction) coarse-grained sense inventories are preferable (Palmer et al., 2004). In these cases, fine-grained distinctions may be unnecessary and even detrimental to WSD and WSI, so that in the last few years many efforts concentrated on clustering senses. Most works focused on producing coarser-grained sense inventories, to the ends of grouping together the closest (partially overlapped) senses of a word; to these ends, various techniques have been carried out, that are briefly surveyed in Section 2.

Differently from existing approaches, we propose a simple yet effective method that relies on recently developed resources that are assumed to also grasp *common-sense knowledge* (Camacho-Collados et al., 2015; Lieto et al., 2016a), which is assumed to be both widely accessible and elementary knowledge (Minsky, 2000), and to reflect *typicality traits* encoded as prototypical knowledge (Rosch, 1975). The research question presently addressed is thus: To what extent can we individuate few principal —common-sense— senses for a term, and in how far is it possible to approximate human performance? Although it is known that even human annotators provide quite different response when annotating text with senses (Palmer et al., 2004), we presently explore the hypothesis that wide-coverage resources are sufficient to individuate the main senses associated to English terms.

2 Related Work

In order to attain coarse-grained senses, different approaches have been proposed, based on some sort of semantic underspecification (Buitelaar, 2000; Ng et al., 2003; Palmer et al., 2007), on existing dictionaries and on exploiting handcrafted sense hierarchies (Navigli, 2006), on syntactic and semantic properties (such as selectional restrictions on verb arguments) (Artale et al., 1998; Palmer et al., 2004), on linguistically motivated heuristics (Mihalcea and Moldovan,

2001), or on distributional similarity among word senses (Agirre and De Lacalle, 2003). Further approaches have been proposed that rely on an adjustable nearest neighbour schema for clustering senses according to the sense granularity actually required by the application at hand (McCarthy, 2006). A popular testbed for experimenting these and other approaches is represented by the sense-annotated *corpora* Senseval-2 and 3 (Edmonds and Cotton, 2001; Mihalcea and Edmonds, 2004).

The problem of annotating a term with the appropriate sense is a challenging one, to such an extent that by no means "two lexicographers working independently are guaranteed to derive the same set of distinctions for a given word" (Palmer et al., 2004). It has been raised that this issue can be overcome to some extent by adopting a more flexible annotation schema, where senses are described in a *graded fashion*: in this way, the applicability of a sense can be assessed on an ordinal scale, rather than in 'crisp' fashion. This sort of annotation would allow to better interpret human annotations, in particular for coarse-grained groups (Erk et al., 2013). A related and complementary issue is that of *clusterability*, that measures in how far word meanings can be partitioned. In this setting, whereas highly clusterable lemmas can be grouped based on traditional clustering techniques, less clusterable lemmas require more sophisticated soft-clustering algorithms to computational systems, and more time and expertise to human annotators (McCarthy et al., 2016).

This work is framed in the context of a long-term project aimed at investigating conceptual categorization (Lieto et al., 2015; Lieto et al., 2016b) based on a hybrid strategy (Evans and Frankish, 2009) complementing formal ontologies with the geometrical framework of Conceptual Spaces (CS) (Gärdenfors, 2014). In particular, we are building a knowledge base to collect conceptual information encoded in a CS-based representational format to provide a uniform interface between the linguistic and the conceptual level, where CSs representations are fully endowed with BabelNet identifiers (Lieto et al., 2016a).[1] This trait will make it possible to link the present work to existing initiatives like Senso Comune (Oltramari and Vetere, 2008; Chiari et al., 2010), that provides about 2000 fundamental Italian terms (De Mauro, 1999) with an ontological description.

3 The CLOSEST Algorithm

The rationale underlying the CLOSEST algorithm is that the main (most frequent) senses gained more room than marginal senses in our lexical and conceptual system and in general in our utterances. This phenomenon determines words and phrases *availability* and *saliency* (Vossen and Fellbaum, 2009), that are arguably grasped by encyclopedic resources, as well. Herein, more central senses are typically featured by richer (i.e., longer vectors) and less specific information, richer semantic connections with other concepts, and heavier feature weights. Although it may happen that some sense spans over (or even subsumes) another one, we are not primarily trying to cluster senses in agglomerative fashion, e.g., by resorting to some superclass of the considered concept; rather, we select the most relevant ones (a term is seldom associated to more than few, say three or four, senses) and we discard the other ones.

The CLOSEST algorithm takes in input a term t and provides a set of possibly related senses.[2] The algorithm first retrieves the set of senses $S = \{s_1, s_2, \ldots, s_n\}$ that are possibly associated to t: such set is obtained by directly querying NASARI. The output of the algorithm is a result set $S^{\trianglelefteq} \subseteq S$. In order to attain $S^{\trianglelefteq}$ we devised a process of incremental filtering, that is arranged into two main phases:

1. *LS-Pruning*. Pruning of less salient senses: senses with associated poor information are eliminated. Senses salience is determined both in absolute terms and in relation to the most salient sense.

2. *OL-Pruning*. Pruning of overlapping senses: if senses with significant overlap are found, the less salient sense is pruned.

Senses are represented as NASARI vectors, that are the vectorial counterpart of BabelNet synsets; concepts (basically, WordNet synsets and Wikipedia pages) are described through vector representations, whose features are synset IDs themselves. Feature weights are computed

[1] The integration of different semantic models such as CSs and the distributional semantics underlying NASARI is still an open issue; we provided an initial solution to this problem in (Lieto et al., 2016a).

[2] The present investigation is restricted to nouns, but no theoretical limitation prevents us from extending the approach to verbs and adjectives.

through the metrics of lexical specificity, by exploiting a semantics-based dimensionality reduction (Camacho-Collados et al., 2015). Each sense is associated with exactly one NASARI vector, so that pruning a sense amounts to pruning a vector.

LS-Pruning. To analyze the senses in S, we inspect each vector $\vec{v}_{ts}$ related to sense s for the term t. The first pruning occurs when no enough information is found, that is when $\vec{v}_{ts}$ contains less than a fixed number of elements (Table 1). Then, in order to determine the next vectors to be pruned, we compute the weight of each vector ($\overline{W}(\vec{v}_{ts})$), the longest vector and the heaviest one among those associated with t ($L(\vec{v}_t)$ and $H(\vec{v}_t)$, respectively). The weight of a NASARI vector $\overline{W}(\vec{v}_{ts})$ is computed by averaging the weight of the features (i.e., the synsets) contained herein. The definitions for these measures are illustrated in Equations 1–3.

$$L(\vec{v}_t) = \arg\max_{s \in S} (\text{len}(\vec{v}_{ts})) \quad (1)$$

$$\overline{W}(\vec{v}_{ts}) = \frac{1}{\text{len}(\vec{v}_{ts})} \cdot \sum_j w_{sj} \quad (2)$$

$$H(\vec{v}_t) = \arg\max_{s \in S} \left(\overline{W}(\vec{v}_{ts})\right). \quad (3)$$

The decision on whether to prune or not a vector is based on a simple criterion: $\vec{v}_{ts} \in S$ is pruned if both its length is below a given fraction of the length of the longest one $L(\vec{v}_t)$, *and* its weight is lower than a given fraction of the heaviest one, $H(\vec{v}_t)$. The parameter settings adopted in the present work are illustrated in Table 1.

OL-Pruning. The second phase of the algorithm aims at detecting overlapped senses. The overlap between vectors that survived the *LS-Pruning* is computed thanks to the information provided in NASARI. The heuristics used in this phase is as follows: the overlap between two vectors $Ovl(\vec{v}_{ti}, \vec{v}_{tj})$ is computed as a fraction of the length of the shortest vector between the two considered, as indicated in Equation 4).

$$Ovl(\vec{v}_{ti}, \vec{v}_{tj}) = \frac{\vec{v}_{ti} \cap \vec{v}_{tj}}{\text{len}(shortest(\vec{v}_{ti}, \vec{v}_{tj}))} \quad (4)$$

The overlapping is checked for every pair $\langle \vec{v}_i, \vec{v}_j \rangle$ (with $i \neq j$) and when an overlap is detected higher than a fixed threshold (see Table 1), the shortest vector between the two is pruned.

At the end of this phase, we have the set $S^{\lhd}$ where only the most salient vectors survived and where, among overlapped vectors, the most salient one has been retained.

3.1 Building the CLOSEST resource

Overall the system handled about $2.69M$ NASARI vectors. Some $207K$ vectors associated to Named Entities were discarded, as not directly related to common-sense concepts; the remaining vectors contained overall $6.9M$ unique words.

The top (most frequent) $15K$ nouns were extracted from the Corpus of Contemporary American English (COCA) which has been built from composite and balanced sources, including spoken, fiction, magazine, newspaper, academic text.[3] Over $6K$ terms were discarded, since they are associated in NASARI either to 1 sense (about $1K$ terms) or to no sense at all (over $5K$ terms), which actually reduced the input size to about $8.7K$ terms; overall $32.6K$ senses were retrieved (on average, 3.7 senses per term), corresponding to such input terms.

The figures featuring the processing phases are reported in Table 1: over $4K$ senses were filtered in the first step of the *LS-Pruning* phase, based on the length of the vector $\vec{v}_{ts}$, and $7.4K$ senses were further discarded in the second step. Finally, in the *OL-Pruning* phase, $5.6K$ vectors were canceled based on overlapping accounts, thus overall yielding $17.5K$ deleted and $15.1K$ survived vectors.[4] The polysemy rate was reduced from the 3.74 senses per term initially featuring NASARI down to 1.73 senses per term, which is in line with the degree of polysemy detected in the Collins English Dictionary for English nouns by WordNet authors (Fellbaum, 1990).

4 Evaluation

A preliminary experimentation has been devised to assess the correctness and completeness of the extracted senses: that is, the question addressed was whether *i)* all senses extracted for the input term are salient (and actually judged as the main senses), and *ii)* all the relevant senses were preserved in CLOSEST. To these ends, 15 volunteers were recruited and interviewed through an on-line questionnaire to evaluate, on a human common-sense judgement basis, the set of senses extracted by the system for 20 terms.

[3] http://corpus.byu.edu/full-text/.

[4] CLOSEST is available at http://goo.gl/7B61Oz.

	condition	threshold values	pruned senses	pruning phase
prune $\vec{v}_{ts}$ IF	$\text{len}(\vec{v}_{ts}) \leq \alpha$	$\alpha = 5$	$4,389$	*LS-Pruning*
	$\left(\frac{\text{len}(\vec{v}_{ts})}{L(\vec{v}_t)} < \beta\right)$ AND $\left(\frac{\overline{W}(\vec{v}_{ts})}{\overline{W}(H(\vec{v}_t))} < \gamma\right)$	$\beta, \gamma = .40$	$7,460$	*LS-Pruning*
	$Ovl(\vec{v}_{ts}, \vec{v}_{tu}) \geq \delta$	$\delta = .20$	$5,676$	*OL-Pruning*

filtered out senses	17,525
retained senses	15,134

Table 1: Pruning of senses in the three steps, along with the number of senses pruned at each step.

Stimuli. The list of 20 terms was algorithmically selected from the aforementioned COCA corpus (see footnote 3) by selecting terms herein with index 1, 51, 101, and so forth. In this way we selected highly frequent terms that are expected to be part of common-sense for those who participated in our experimentation.[5]

Experimental design and procedure. The participants were asked *a)* to assess each and every sense extracted by the system and associated to each input term by indicating whether it was acceptable as one of the principal senses for the term at hand. Additionally, they were requested *b)* to indicate any further sense they reputed essential in order to complete the common-sense pool of senses for the given term.

Results. Overall 42 senses (corresponding to the 20 mentioned terms) were assessed through the experimentation: each sense was rated 15 times, thus resulting in 630 judgements: 24% of senses were not found appropriate, according to a common-sense judgement, thereby determining a 76% accuracy as regards as question *a)*. However, if we consider senses refused by at least 10 participants, only 5 senses were refused (12%), that actually correspond to very specific senses (e.g., the sense 'Net (textile)' for the term 'network'; 'Session (Presbyterianism)', 'session house' for the term 'session').

As regards as question *b)*, results are more difficult to interpret, due to the sparsity of the answers: out of the 59 added senses, only in 8 cases the added sense has been indicated by two or three participants (and never more): in such cases it emerged, for example, that the sense 'manners' was relevant (and missing, in the CLOSEST resource) for the input term 'education'; the sense 'social network' is relevant for the term 'network'; and 'meeting' for 'session'.

However, although encouraging results emerged from the experimentation, further experiments are needed to assess the CLOSEST resource in a more extensive and principled way, also in consideration of the many factors that were presently neglected, such as, e.g., age, education, occupation of the participants, their native language, *etc.*.

5 Conclusions

In this paper we have illustrated the CLOSEST algorithm to extract the most salient (under the common-sense perspective) senses associated to a given term; also, we have introduced the CLOSEST resource, which has been built by starting from the $15K$ top frequency English terms. The resource currently provides senses in a flat manner, but, if required, senses can be organized in a sorted fashion by extending the metrics used for filtering. Our work relies on a recently developed resource such as NASARI that is multilingual in nature.[6] Consequently, different from most previous approaches, CLOSEST can be linked to various existing resources aimed at grasping common-sense to complete the ideal chain connecting lexicon, semantics and formal (ontological) description. The experimentation revealed a reasonable agreement with human responses, and pointed out some difficulties in fully assessing this sort of resource. These issues, along with improvements to the heuristics implemented by the algorithm and a different evaluation based on a shared NLP task, will be addressed in future work.

[5] The full list of the considered terms includes: time, side, education, type, officer, ability, network, shoulder, threat, investigation, gold, claim, learning, session, aid, emergency, bowl, pepper, milk, resistance. The printed version of the online questionnaire is available at the URL `http://goo.gl/w9TNQT`.

[6] An interesting question may be raised on this point, about the conceptual alignment in a *inter*-linguistic perspective, which is a well-known issue, e.g., for applications in the legal field (Ajani et al., 2010).

References

Eneko Agirre and Oier Lopez De Lacalle. 2003. Clustering WordNet Word Senses. In *RANLP*, volume 260, pages 121–130.

Gianmaria Ajani, Guido Boella, Leonardo Lesmo, Marco Martin, Alessandro Mazzei, Daniele P Radicioni, and Piercarlo Rossi. 2010. Multilevel legal ontologies. In *Semantic Processing of Legal Texts*, pages 136–154. Springer.

Alessandro Artale, Anna Goy, Bernardo Magnini, Emanuele Pianta, and Carlo Strapparava. 1998. Coping with WordNet Sense Proliferation. In *First International Conference on Language Resources & Evaluation*.

Paul Buitelaar. 2000. Reducing Lexical Semantic Complexity with Systematic Polysemous Classes and Underspecification. In *NAACL-ANLP 2000 Workshop: Syntactic and Semantic Complexity in Natural Language Processing Systems*, pages 14–19. Association for Computational Linguistics.

José Camacho-Collados, Mohammad Taher Pilehvar, and Roberto Navigli. 2015. NASARI: a Novel Approach to a Semantically-Aware Representation of Items. In *Proceedings of NAACL*, pages 567–577.

Isabella Chiari, Alessandro Oltramari, and Guido Vetere. 2010. Di Cosa Parliamo quando Parliamo Fondamentale? Lessemi, Accezioni, Sensi e Ontologie. In *Lessico e Lessicologia. Atti del Convegno della Societ di Linguistica Italiana*, pages 177–194, Roma, September. Bulzoni.

Tullio De Mauro. 1999. *Grande Dizionario Italiano dell'Uso*. UTET, Turin, Italy.

Philip Edmonds and Scott Cotton. 2001. SENSEVAL-2: Overview. In *Proceedings of SENSEVAL-2 Second International Workshop on Evaluating Word Sense Disambiguation Systems*, pages 1–5, Toulouse, France, July. Association for Computational Linguistics.

Katrin Erk, Diana McCarthy, and Nicholas Gaylord. 2013. Measuring word meaning in context. *Computational Linguistics*, 39(3):511–554.

Jonathan St BT Evans and Keith Ed Frankish. 2009. *In Two Minds: Dual Processes and Beyond*. Oxford University Press.

Christiane Fellbaum. 1990. English Verbs as a Semantic Net. *International Journal of Lexicography*, 3(4):278–301.

Peter Gärdenfors. 2014. *The Geometry of Meaning: Semantics Based on Conceptual Spaces*. MIT Press.

Antonio Lieto, Daniele P. Radicioni, and Valentina Rho. 2015. A Common-Sense Conceptual Categorization System Integrating Heterogeneous Proxytypes and the Dual Process of Reasoning. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, pages 875–881, Buenos Aires, July. AAAI Press.

Antonio Lieto, Enrico Mensa, and Daniele P. Radicioni. 2016a. A Resource-Driven Approach for Anchoring Linguistic Resources to Conceptual Spaces. In *Proceedings of the 15th International Conference of the Italian Association for Artificial Intelligence*, Genoa, Italy, December. Springer.

Antonio Lieto, Daniele P Radicioni, and Valentina Rho. 2016b. Dual PECCS: a Cognitive System for Conceptual Representation and Categorization. *Journal of Experimental & Theoretical Artificial Intelligence*, pages 1–20.

Diana McCarthy, Marianna Apidianaki, and Katrin Erk. 2016. Word Sense Clustering and Clusterability. *Computational Linguistics*.

Diana McCarthy. 2006. Relating WordNet Senses for Word Sense Disambiguation. *Making Sense of Sense: Bringing Psycholinguistics and Computational Linguistics Together*, 17.

Rada Mihalcea and Phil Edmonds. 2004. SENSEVAL-3: Overview. In *Proceedings Senseval-3 3rd International Workshop on Evaluating Word Sense Disambiguation Systems. ACL, Barcelona, Spain*.

Rada Mihalcea and Dan I Moldovan. 2001. Automatic Generation of a Coarse Grained WordNet. In *Proceedings of the NAACL Workshop on WordNet and Other Lexical Resources*.

George A Miller. 1995. WordNet: a Lexical Database for English. *Communications of the ACM*, 38(11):39–41.

Marvin Minsky. 2000. Commonsense-based interfaces. *Communications of the ACM*, 43(8):66–73.

Roberto Navigli and Simone Paolo Ponzetto. 2010. BabelNet: Building a Very Large Multilingual Semantic Network. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 216–225. Association for Computational Linguistics.

Roberto Navigli. 2006. Meaningful Clustering of Senses Helps Boost Word Sense Disambiguation Performance. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 105–112. Association for Computational Linguistics.

Hwee Tou Ng, Bin Wang, and Yee Seng Chan. 2003. Exploiting Parallel Texts for Word Sense Disambiguation: An Empirical Study. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pages 455–462. Association for Computational Linguistics.

Alessandro Oltramari and Guido Vetere. 2008. Lexicon and Ontology Interplay in Senso Comune. *OntoLex 2008 Programme*, page 24.

Martha Palmer, Olga Babko-Malaya, and Hoa Trang Dang. 2004. Different Sense Granularities for Different Applications. In *Proceedings of Workshop on Scalable Natural Language Understanding*.

Martha Palmer, Hoa Trang Dang, and Christiane Fellbaum. 2007. Making Fine-Grained and Coarse-Grained Sense Distinctions, both Manually and Automatically. *Natural Language Engineering*, 13(02):137–163.

Eleanor Rosch. 1975. Cognitive Representations of Semantic Categories. *Journal of Experimental Psychology: General*, 104(3):192–233.

Piek Vossen and Christiane Fellbaum, 2009. *Multilingual FrameNets in Computational Lexicography: Methods and Applications*, chapter Universals and idiosyncrasies in multilingual WordNets. Trends in linguistics / Studies and monographs: Studies and monographs. Mouton de Gruyter.