CSI: A Coarse Sense Inventory for 85% Word Sense Disambiguation

Caterina Lacerra*, **Michele Bevilacqua***, **Tommaso Pasini**, **Roberto Navigli**
Sapienza University of Rome
Department of Computer Science
{lacerra, bevilacqua, pasini, navigli}@di.uniroma1.it

Abstract

Word Sense Disambiguation (WSD) is the task of associating a word in context with one of its meanings. While many works in the past have focused on raising the state of the art, none has even come close to achieving an F-score in the 80% ballpark when using WordNet as its sense inventory. We contend that one of the main reasons for this failure is the excessively fine granularity of this inventory, resulting in senses that are hard to differentiate between, even for an experienced human annotator. In this paper we cope with this long-standing problem by introducing Coarse Sense Inventory (CSI), obtained by linking WordNet concepts to a new set of 45 labels. The results show that the coarse granularity of CSI leads a WSD model to achieve 85.9% F1, while maintaining a high expressive power. Our set of labels also exhibits ease of use in tagging and a descriptiveness that other coarse inventories lack, as demonstrated in two annotation tasks which we performed. Moreover, a few-shot evaluation proves that the class-based nature of CSI allows the model to generalise over unseen or under-represented words.

1 Introduction

Word Sense Disambiguation (WSD) is the task of assigning the correct meaning from among a finite set of possible choices to a word in a context (Navigli 2009). It is a key task in Natural Language Processing (Navigli 2018), providing semantic information that is potentially beneficial for downstream applications, such as information extraction (Delli Bovi, Espinosa Anke, and Navigli 2015) and machine translation (Pu et al. 2018). While much effort has been devoted to building new algorithms or data (Pasini and Navigli 2018; Scarlini, Pasini, and Navigli 2019) for this task, state-of-the-art systems have yet to break the 80% accuracy ceiling on standard WSD benchmark datasets (Raganato, Delli Bovi, and Navigli 2017; Bevilacqua and Navigli 2019; Vial, Lecouteux, and Schwab 2019; Scarlini, Pasini, and Navigli 2020), showing that the WSD task is far from being solved. Following the literature in the field (Hovy et al. 2006; Palmer, Dang, and Fellbaum 2007; Navigli, Litkowski, and Hargraves 2007), we argue that the reason for this unsatisfactory performance does not lie solely in the complexity of the task but also in the fine granularity of the sense inventory adopted, i.e., WordNet (Fellbaum 1998). For example the noun *street* has separate WordNet senses for the 'thoroughfare (usually including sidewalks)' and 'the part of a thoroughfare between the sidewalks'. Such fine-grained distinctions introduce noise and sparsity for machine learning algorithms in a task where reliable data is very costly to produce. Moreover, the inter-annotator agreement with WordNet ranges from 0.6 to 0.8 (Navigli 2009), making it clear that, unless super-human performance is expected, WSD systems will not exceed this ceiling. To overcome these issues, in this paper we present Coarse Sense Inventory (CSI), a new organization of concepts based on a large-scale mapping of WordNet synsets to domain-based semantic labels. CSI labels are tailored to WSD, with each label shared across different words and part-of-speech (POS) tags. The inventory has been developed starting from the categories of a general domain thesaurus, i.e., Roget's (2011), which have been clustered into coarser labels, leading to an inventory whose high-level semantics is domain-based (describing what each label is *about*) rather than hypernymy-based (what the label *is a kind of*). The experimental results provide evidence that CSI is better suited for WSD than other existing competitors; moreover, CSI leads a supervised neural system to reach an F1 of almost 86% overall and helps the model to generalise over unseen or under-represented words.

In this work, we provide four main contributions:

1. We introduce CSI, a new coarse-grained sense inventory where semantic labels are shared across the lexicon.
2. Our new sense inventory achieves better qualitative results on two manual annotation tasks than its alternatives: our labels enable a higher inter-annotator agreement and are more descriptive than those of the competitors.
3. CSI outperforms all the compared coarse-grained sense inventories on all-words WSD, attaining a better trade-off between performance and expressiveness.
4. CSI paves the way to few-shot learning in WSD, as it reaches better performances than its alternative inventories on unseen and under-represented words.

*The authors contributed equally.

2 Related Work

A sense inventory enumerates the possible meanings that content words (nouns, verbs, adverbs, adjectives) may assume. Even though enumerative representations of lexical semantics have been the object of some criticism (Kilgarriff 1997), the enumerative lexicon is still the most popular approach in WSD as it defines a possible finite ground truth for word meanings. Indeed, WordNet is the *de facto* standard sense inventory for WSD, with it being the largest manually-crafted and freely available inventory, grouping 155287 different lemmas (word form-POS pairs) in 117659 concepts called synsets, i.e., sets of synonyms (statistics for version 3.0). The main criticism that is made against WordNet is that its fine granularity and subtle distinctions between nearly identical senses make it hard to select the most suitable meaning of a given word, even for humans (Edmonds and Kilgarriff 2002). To overcome this problem, different sense inventories with coarser granularity have been developed. Following Izquierdo, Suarez, and Rigau (2015), we group inventories into two categories: i) *word-based* and ii) *class-based*. We review these two groups in what follows.

Word-based Many works in the past have proposed different approaches for solving the fine-granularity problem with coarser sense inventories created by clustering WordNet senses that are associated with the same lemma (Palmer, Babko-Malaya, and Dang 2004; Palmer, Dang, and Fellbaum 2007). Hovy et al. (2006) introduced the OntoNotes project, whose objectives included the release of a manually-built sense inventory. The resource was obtained by iteratively merging senses until 90% inter-annotator agreement was reached, thus encouraging annotators to choose coarser senses to meet the agreement goal. Differently, but pursuing the same objective, the work of Navigli (2006) coarsened the WordNet inventory by clustering and mapping its word senses to the Oxford Dictionary of English (ODE). This work was later used as the starting point for introducing the task of coarse-grained all-words WSD in the context of SemEval-07 (Navigli, Litkowski, and Hargraves 2007). With the same purpose of reducing the granularity of senses, Snow et al. (2007) proposed a supervised approach to predict whether two senses should be merged or not.

Class-based One of the drawbacks of word-based approaches is that their sense labels are still tied to words, thus leaving unsolved the problem of rare senses which have none, or only few, occurrences in an annotated corpus. Class-based approaches, instead, cope with this issue by providing labels that are shared among different words, enabling a more efficient usage of annotated data, and mitigating the problem of the long tail of infrequent word senses.

One of the earliest approaches treats WordNet's lexicographer files as coarse classes, which we refer to as SuperSenses: each class includes synsets with the same part of speech and a broad semantic type, like VERB.PERCEPTION. While each WordNet synset is associated with one label from a set of 45 available, the 4 labels used for adverbs and adjectives are not semantically meaningful, making the resource of limited usefulness for all-words WSD. For example, all the senses of the adjective *bright* are classified as ADJ.ALL, making it impossible to distinguish the *intelligent* meaning of the word from the *light* one. Izquierdo, Suárez, and Rigau (2007), instead, exploited WordNet relations to automatically extract a set of fundamental senses, called Basic Level Concepts, to which all the other senses are mapped. Similarly, Vial, Lecouteux, and Schwab (2019) leveraged hypernymy to reduce the WordNet granularity, releasing a 39K-label inventory used for fine-grained WSD. Another WordNet-based resource is WordNet Domains (Magnini and Cavaglià 2000), a mapping from WordNet synsets to a set of 200 labels loosely following the Dewey Decimal Classification system (Dewey 1876). The authors took a semi-supervised approach where they manually annotated a moderate number of seed synsets and then propagated the labels by exploiting the WordNet structure. Along the lines of WordNet Domains, Camacho-Collados and Navigli (2017) introduced BabelDomains, a set of 42 labels grouping in coarser-grained classes the nominal synsets of BabelNet (Navigli and Ponzetto 2012), a multilingual knowledge base comprising WordNet, Wikipedia and other resources. BabelDomains employs top-level categories from Wikipedia featured articles, thus making it able to cover a comprehensive set of knowledge domains.

Differently from the aforementioned class-based inventories, CSI has been manually created from scratch. Moreover, in contrast to BabelDomains – which covers only nouns – and SuperSenses and Basic Level Concepts – in which only nouns and verbs are meaningfully clustered – CSI covers all the content-word POS tags. Furthermore, our coarse inventory encompasses semantic areas that are excluded from other existing resources, inter alia, the five senses of perception and the areas of routines and daily activities.

3 CSI: A Coarse Sense Inventory

In this Section we present our novel class-based Coarse Sense Inventory (CSI). The main objective of our approach is to build a resource that avoids sense distinctions that are too fine-grained for WSD, while maintaining a granularity that is still meaningful for the task. Our method (see Figure 1) consists of two steps: i) *tagset definition*, in which, by clustering Roget's categories, we define a new coarse-grained sense inventory, and ii) *synset mapping*, where we map WordNet synsets to one or more CSI labels.

Tagset Definition The first step aims at building a set of coarse labels which covers the largest possible portion of the semantic space. For this purpose we exploit Roget's thesaurus, a widely-used resource in NLP which provides a categorization for the lexicon of the English language. The thesaurus contains 1075 categories, grouped into 15 broader classes, none of which offers the level of granularity we need for the purpose of class-based WSD. In fact, categories under the same class may be either too similar, such as *tribunal, jury, lawyer*, or too different, such as *lawyer* and *learning*. At the same time, the 15 Roget's classes are not specific

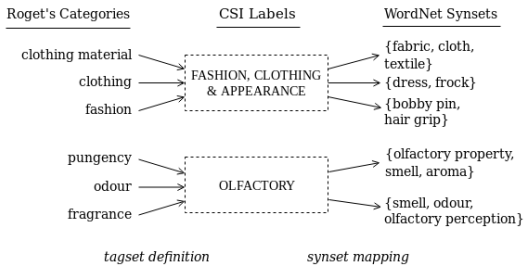

Figure 1: An excerpt of the mapping between Roget's categories, CSI labels and WordNet synsets.

enough to be used as labels, since they describe broad domains such as *Values and the ideals*. For this reason we asked three expert linguists to group the Roget's categories into clusters which could serve as labels for our sense inventory. For each of the Roget's classes the annotators identified those categories representing semantically unrelated fields that could potentially appear in different contexts (e.g., *hair, sleep* and *color* in the class *the body and the senses*). When this was the case, they could either create one or more new clusters, or assign such categories to an existing cluster. Finally, once the taggers converged on a common set of clusters, they named each of them by considering the shared semantics of the categories and the possible application context they could appear in.

For example, the *jurisdiction, tribunal, judge, jury* and *lawyer* categories were merged into a single cluster named LAW&CRIME, while the *fragrance* and *odor* categories were grouped together under the OLFACTORY label. Moreover, we also added the semantically-empty label named GENERAL to our inventory to cover the 135 categories that could not be included in any cluster. As a result, CSI covers all the semantic areas expressed by the 1075 Roget's categories. Note that some categories can belong to more than one cluster, leading to partitions of the semantic space that are not disjoint. For example, the category *religious buildings* belongs both to the cluster named ART,ARCHITECTURE&ARCHAEOLOGY and to the cluster RELIGION,MYSTICISM&MYTHOLOGY.

Synset Mapping The second step aims at mapping the WordNet synsets to one or more CSI labels. To this end, the annotators iterated over each WordNet synset, associating it with one or more CSI labels, exploiting as context its gloss and its occurrences in the sentences of SemCor (Miller et al. 1993), i.e., a manually-annotated corpus. The annotators mapped a sample of WordNet composed of the most frequent synsets occurring in SemCor, so as to guarantee a large coverage of its instances. The outcome of this step is a mapping of 8217 synsets to one or more coarse labels, covering 78% of the annotated instances in SemCor. For example, the synsets {*plebeian, pleb*} and {*Marxism*} are associated with the CSI label POLITICS,GOVERNMENT&NOBILITY, while {*treaty, pact, accord*} is labelled with both LAW&CRIME and POLITICS,GOVERNMENT&NOBILITY.

To further increase CSI coverage, we mapped each of the labels of BabelDomains (Camacho-Collados and Navigli 2017) to one or more CSI tags. To ensure the consistency of the labels, an annotator manually validated all the CSI labels that did not have a one-to-one correspondence with BabelDomains tags. This mapping guarantees an additional coverage of 78K WordNet synsets. As a result, a total of 83K synsets were annotated with at least one coarse label, encompassing all open-class parts of speech.

4 Experiments

We now present a set of experiments aimed at assessing the quality of CSI under different perspectives: i) we designed two annotation tasks to evaluate the reliability (Section 4.1) and the descriptiveness (Section 4.2) of the labels in each coarse-grained inventory, ii) we exploited the WSD task to evaluate and compare CSI with other class-based inventories, and iii) we tested CSI's ability to enable zero- and few-shot learning, also in comparison with its alternatives.

Competitors As competitors of CSI, we considered a word-based and fine-grained inventory, i.e., WordNet, and three class-based and coarse-grained inventories, i.e., BabelDomains (BD), WordNet Domains (WND) and SuperSenses (SuS). As regards the class-based ones, we recall from Section 2 that they provide a mapping from the WordNet synsets to one or more of their coarse labels. Among our comparisons we did not calculate the improvements brought by CSI with respect to a random clustering, as proposed by Snow et al. (2007), since we were mainly interested in evaluating the use of our labels in coarse-grained WSD rather than the clustering itself, and because the proposed metric can only be applied to disjoint clusters, which was not our case.

4.1 Label Selection

To test whether the use of CSI can result in more reliable annotations, we designed a task where three expert linguists - not involved in the creation of the mapping - were asked to annotate 200 target words according to the inventories under comparison. For each coarse inventory we defined the set of possible labels for a given word as the union of the labels associated with the word's synsets. For WordNet, we directly considered the glosses of the word's meanings as labels, instead. Then, for each target word, we provided the annotators with a context sentence from SemCor, together with the set of possible labels for that word in each inventory. For example, we presented the following sentence to the annotators: "Madden *settled* back to read the will" and they had to choose among the CSI, WordNet, SuperSenses and WordNet Domains senses of *settled*, that are, respectively, {SPACE&TOUCH, BUSINESS,ECONOMICS&FINANCE, LAW&CRIME}, {*settle into a position*;*bring to an end*; ...; *end a legal dispute*}, {VERB.CHANGE, VERB.COGNITION}, and {POLITICS, FACTOTUM}.

Measures To evaluate the inter-annotator agreement we calculated the Kraemer's κ coefficient (Kraemer 1980). We preferred it to the better known Cohen's κ (Cohen 1960), of

which the former is an extension, since it allows the annotators to provide more than one answer for an item. To compute the Kraemer's κ we first represent the response A_i^j of each annotator j for an item i as a vector $r_i^j \in \mathbb{R}^{|L|}$, where L is the set of possible labels. Within r_i^j, each dimension corresponds to one of the labels $l_1, \ldots, l_{|L|} \in L$ and can take one of the following two values: the mean of $1, \ldots, |A_i^j|$ for dimensions corresponding to labels $\in A_i^j$ and the mean of $|A_i^j|+1, \ldots, |L|$ for the others. Formally:

$$r_i^j[k] = \begin{cases} \frac{1}{|A_i^j|} \sum_{n=1}^{|A_i^j|} n & \text{if } l_k \in A_i^j \\ \frac{1}{|L|-|A_i^j|} \sum_{n=|A_i^j|+1}^{|L|} n & \text{otherwise} \end{cases}$$

where $k \in \{1, \ldots, |L|\}$. For example, supposing annotator 2 gave the CSI response labels $A_i^2 = \{$BIOLOGY, CHEMISTRY&MINERALOGY$\}$ for item i, we calculate a value of $\frac{1}{2}\sum_{n=1}^{2} n = 1.5$ for the labels chosen by the annotator and a value of $\frac{1}{45}\sum_{n=3}^{45} n = 23$ for all the others[1]. Assuming that BIOLOGY and CHEMISTRY&MINERALOGY correspond to the indices 2 and 4, we build the following rank vector $r_i^2 = (23, 1.5, 23, 1.5, \ldots, 23)$. Once each annotation A_i^j has its associated rank vector r_i^j, we can proceed to compute their correlation. To do so, we calculate the mean R_I of Spearman's correlations between all pairs of rank vectors r_i^j and r_i^k for each item i, i.e., $R_I = N^{-1}\sum_{i=1}^{N} \rho(r_i^1, r_i^2)$, where N is the number of annotation items. R_I acts as a measure of *observed agreement*, therefore we now need to quantify the agreement by chance. For this purpose we define U as the set comprising all the annotations and compute the Spearman's correlation average between all the annotation pairs in $U \times U$ as follows:

$$R_T = \frac{1}{|U|^2} \sum_{(A_i, A_j) \in U \times U} \rho(r_i, r_j)$$

where A_i and A_j are two annotations in U, r_i and r_j are their corresponding rank vectors and $\rho(r_i, r_j)$ is their Spearman correlation. Finally, Kraemer's κ is calculated as the ratio of the difference between observed agreement (R_I) and chance agreement (R_T), and the difference between perfect agreement and chance agreement: $\kappa = (R_I - R_T)(1 - R_T)^{-1}$. To interpret κ values we followed Landis and Koch (1977), that define the $(0.4, 0.6]$ interval as *moderate agreement*, $(0.6, 0.8]$ as *substantial agreement* and $(0.8, 1.0)$ as *almost perfect agreement*. In computational linguistics, there is a consensus that puts the cutoff above which the annotations are considered reliable at 0.67 (Di Eugenio and Glass 2004).

Results In Table 1 (first row) we report the Kraemer's κ coefficient attained with CSI, WordNet Domains (WND), SuperSenses and WordNet. As one can see, when the annotations are carried out with CSI, the annotators tend to agree more than when using other coarse-grained or fine-grained sense inventories. Indeed, the agreement achieved when using CSI falls in the *almost perfect* part of the spectrum of κ values according to the literature (Landis and Koch 1977), while, when using SuperSenses and WordNet Domains inventories, the agreement has to be considered *substantially reliable*. As expected, instead, due to their fine granularity, WordNet senses allow only a *moderate agreement*, hence confirming the results reported by Palmer, Dang, and Fellbaum (2007). These outcomes show that CSI labels provide useful semantic information that make them easier to use than the labels of the other sense inventories, hence simplifying the task of annotating large amounts of data.

[1] We recall from Section 3 that CSI has 45 labels, i.e., $|L| = 45$.

Measure	CSI	WND	SuperSenses	WordNet
IAA	**0.81**	0.74	0.69	0.51
Descriptiveness	**2.23**	1.80	2.04	-

Table 1: Kraemer's κ agreement (first row); average descriptiveness for coarse inventories' labels (second row).

Sentence	The street that is full now of traffic and parked **cars** drowsed on an August afternoon in the **shade** of the curbside trees, and **silence** was a weight [...].		
Word	CSI	SuperSenses	WND
cars	TRANSPORT&TRAVEL	N.ARTIFACT	TOURISM
shade	PHYSICS&ASTRONOMY	N.STATE	FACTOTUM
silence	MUSIC, SOUND&DANCING	N.ATTRIBUTE	ACOUSTICS

Table 2: An example of how target words from SemCor are annotated in CSI, SuperSenses and WordNet Domains.

4.2 Descriptiveness

In this Section we assess the extent to which CSI and the other coarse-grained inventories provide labels that are easy to understand for humans. Specifically, we are interested in studying the degree of both pertinence and informativeness that characterise each inventory when using it to tag a text. To this end, we designed an annotation task for 150 words in which, given a target word in a sentence from SemCor and its gold label according to each of the coarse inventories, the three annotators had to rank the labels in increasing order of descriptiveness for the given target word. For example, as shown in Table 2, we presented the annotators with the four target words *street*, *cars*, *shade* and *silence*, together with their corresponding sentence "the street that is full [. . .]", and asked them to rank the three labels provided for each word by CSI, SuperSenses and WordNet Domains. As an annotation for a given target word, the linguists were asked to provide a score ranking ranging from 1 for the labels that were less descriptive to 3 for those that were the most descriptive (ties were allowed).

Measures To evaluate the descriptiveness of each inventory under comparison, we calculated the average rank of the labels across all the 150 annotations. Formally,

$$descriptiveness(I) = \frac{1}{|N||J|} \sum_{j \in J} \sum_{(t,s) \in N} rank_j(l_I^{(t,s)})$$

Coverage	SC instances		SC synsets		WN synsets	
	total	%	total	%	total	%
CSI	198K	88.0	16K	61.7	83K	70.4
SuperSenses	226K	100.0	26K	100.0	118K	100.0
WND	163K	72.1	18K	69.5	93K	78.7
Intersection	153K	67.5	14K	54.0	79K	67.1

Table 3: Coverage of SemCor (SC) and WordNet (WN) by class-based inventories.

where J is the set of annotators, $l_I^{(t,s)}$ is the label with which the target word t in the sentence s is tagged according to inventory I, $rank_j(x)$ is the rank given by annotator j to x and N is the set of annotations to be carried out.

Results We now report the scores attained in the descriptiveness task, computed over the responses provided by the three annotators. As can be seen in the second row of Table 1, CSI is the inventory with the highest scores, proving on average to be the one with the most descriptive labels, while those of WordNet Domains and SuperSenses are ranked lower, meaning that they do not offer an adequate degree of characterization. In fact, considering the example in Table 2, the labels provided by both SuperSenses and WND are either inappropriate, i.e., the WND TOURISM label associated with *car*, or too general and hence not informative, i.e., the FACTOTUM label of WND for *shade* and the SuperSenses label ATTRIBUTE for *silence*. CSI, instead, provides a more precise and detailed information on each word in bold in the example. In summary, not only did CSI prove to be the inventory with the highest ease of use compared to its competitors (see Section 4.1), but it also exhibited a higher degree of descriptiveness in its labels, which can increase the readability of a text, i.e., making it easier for humans to understand it. It is reasonable to expect, in fact, that very descriptive labels, such as those provided in CSI, might also be useful on their own, e.g., to improve the reading comprehension of a language learner.

4.3 All-Words Word Sense Disambiguation

In this Section we compare CSI with the aforementioned sense inventories (see the beginning of Section 4), by using them as labels for a WSD system.

WSD Models To evaluate the performances of our inventory across different learning models, we implemented two neural WSD systems. The first model was similar to that described in Vial, Lecouteux, and Schwab (2019), featuring two bidirectional LSTM layers, that encode the context of each token, and an attention mechanism. The two vectors (the attention and LSTM outputs) are concatenated and fed into a dense layer for classification. As input features, we tried two different pre-trained contextual embeddings, i.e., ELMo (Peters et al. 2018) and BERT base (Devlin et al. 2019). The second model, instead, had a simpler architecture in which BERT large contextualized embeddings are fed directly to a fully-connected layer for classification.

Data For training, developing and testing we used the WSD evaluation framework made available by Raganato, Camacho-Collados, and Navigli (2017). It includes SemCor, which we used as training set, and the 5 standard all-words WSD benchmarks for English, i.e., Senseval-2 (Palmer et al. 2001), Senseval-3 (Snyder and Palmer 2004), SemEval-07 (Pradhan et al. 2007), SemEval-13 (Navigli, Jurgens, and Vannella 2013), SemEval-15 (Moro and Navigli 2015). We will use ALL to refer to the concatenation of all the foregoing benchmark datasets except SemEval-07, which, following Raganato, Camacho-Collados, and Navigli (2017), we used as development set.

In order to evaluate each sense inventory we replaced the WordNet sense keys appearing in the training set and in all test sets with their coarse labels in the sense inventory under assessment. Whenever a sense key in the training set occurred that was associated with multiple labels, we replaced it with a random tag taken from among the mapped ones. For the evaluation, we considered the predicted label to be correct if it was included in the set of all the possible labels for the instance. Finally, to set a level playing field among all the sense inventories under comparison, we considered only the training and testing instances that were in common, i.e., all those tagged with a synset that was covered by each of the inventories. Therefore, to consider the intersection, we restricted the training data to 153K out of 226K instances, as shown in the last row of Table 3, where we also report the coverage of each inventory with respect to the synsets of WordNet and to those that appear in SemCor.

Evaluation Measure As evaluation measure for performance we used F1. However, we note that each inventory makes the task either more or less difficult as there are significant differences in the number of labels and in the way the synsets are grouped. Therefore, to estimate the difficulty of the disambiguation task according to the sense inventory, we computed the perplexity of a random guessing model as the inverse of the probability of choosing the correct answer, by randomly sampling one label from those possible for the given item: $PPL = \frac{1}{m}\sum_{i=1}^{m}\frac{L_{l_i}}{G_i}$, where L_{l_i} is the number of all the labels that are associated with the lemma l_i in a given sense inventory, G_i is the number of correct answers for the instance i and m is the number of instances in the dataset. Since the difficulty of the task (PPL) and the performance of the model (F1) are inversely correlated, we compute a Geometric Trade-Off (GTO) measure, which is the best choice when averaging values with different magnitude (see, e.g., Komninos and Manandhar (2016)).

Hyperparameters Every model variant freezes the embedding layer's weights during training, and the input to the network is fed in batches of 64. In the first model, the output of the two Bi-LSTMs layers is set to 512. When using ELMo, the sentences longer than 30 words are truncated. As optimizer we used Adam (Kingma and Ba 2015) with learning rate 10^{-3} and 10^{-4} for ELMo and BERT, respectively.

		F1						PPL						GTO					
	Inventory	SE-2	SE-3	SE-07	SE-13	SE-15	ALL	SE-2	SE-3	SE-07	SE-13	SE-15	ALL	SE-2	SE-3	SE-07	SE-13	SE-15	ALL
ELMo	CSI	83.5	81.7	79.9	81.9	77.9	81.7	2.62	3.13	3.71	2.28	2.93	**2.70**	1.48	1.6	1.72	1.37	1.51	**1.49**
+	WND	89.8	86.5	91.7	80.6	85.0	**85.5**	2.00	2.33	2.25	2.12	2.01	2.13	1.34	1.42	1.44	1.31	1.31	1.35
LSTM	SuS	82.3	78.9	81.5	79.8	80.2	80.3	2.25	2.69	2.98	2.15	2.26	2.34	1.36	1.46	1.56	1.31	1.34	1.37
BERT	CSI	84.8	83.4	75.7	80.3	76.8	81.9	2.62	3.13	3.71	2.28	2.93	**2.70**	1.49	1.62	1.67	1.35	1.50	**1.49**
+	WND	87.4	85.3	89.1	82.1	81.0	**84.4**	2.00	2.33	2.25	2.12	2.01	2.13	1.32	1.41	1.42	1.32	1.28	1.34
LSTM	SuS	81.5	79.1	79.4	79.0	79.6	79.8	2.25	2.69	2.98	2.15	2.26	2.34	1.36	1.46	1.54	1.30	1.34	1.37
BERT	CSI	86.0	84.5	75.5	83.3	79.3	83.8	2.62	3.13	3.71	2.28	2.93	**2.70**	1.50	1.63	1.67	1.38	1.53	**1.51**
+	WND	91.2	87.9	89.4	83.7	84.7	**87.2**	2.00	2.33	2.25	2.12	2.01	2.13	1.35	1.43	1.42	1.33	1.30	1.36
Dense	SuS	83.2	81.4	79.9	80.5	82.3	81.8	2.25	2.69	2.98	2.15	2.26	2.34	1.37	1.48	1.54	1.32	1.36	1.38

Table 4: Comparison of CSI against WordNet Domains (WND) and SuperSenses (SuS) on all-words WSD tasks from past Senseval and SemEval competitions.

CSI vs. BabelDomains						
Model	F1		Perplexity		GTO	
	CSI	BD	CSI	BD	CSI	BD
ELMo + Bi-LSTM	86.9	86.9	1.92	1.95	1.29	**1.30**
BERT + Bi-LSTM	89.4	87.1	1.92	1.95	**1.31**	1.30
BERT + Dense	91.1	89.1	1.92	1.95	**1.32**	**1.32**

Table 5: Comparison of CSI against BabelDomains (BD), on all-words WSD. Results are shown for the ALL dataset.

Results We now report the results of the comparison between CSI and its competitors. For each sense inventory we computed its performance in terms of F1, the perplexity of random guessing with the considered inventory and their geometric mean. We remark that a higher perplexity indicates a higher uncertainty of random guessing, i.e., the sense inventory associates on average a higher number of possible labels with a given lemma, thus making the disambiguation task harder. Since CSI extends BabelDomains, we first report their comparison in Table 5. As can be seen, the two inventories reach a GTO score that is almost the same across the WSD models. However, BabelDomains is inherently limited for WSD tasks as it only covers nouns, while CSI covers all the open-class parts of speech without losing anything in terms of performance compared to BabelDomains. Since CSI proved to be on a par with BabelDomains while, at the same time, having a wider coverage of POS tags, in what follows we only report the results for CSI. We now move to compare our inventory with the other class-based approaches, i.e., WordNet Domains and SuperSenses. As shown in Table 4, CSI consistently attains better GTO scores than the other inventories, proving its better balance between label granularity and expressiveness, regardless of the underlying neural model. More in detail, we note that, while WordNet Domains achieves higher F1 scores across datasets, except for SemEval-13 with the ELMo + Bi-LSTM model, its perplexity is always lower, meaning that the disambiguation task becomes easier due to the lower expressiveness of the inventory. Indeed, more than 18% of the WordNet synsets are mapped to the semantically-empty label of WND, i.e., FACTOTUM. CSI, in contrast, resorts to GENERAL for less than 1% of the annotated synsets.

Differently from WordNet Domains, SuperSenses achieves F1 scores that are lower than CSI, except for SemEval-07 and SemEval-15. Moreover, it shows a lower perplexity overall, proving to be a less expressive inventory. In fact, it provides only 4 possible classes in total for adjectives and adverbs, thus making the task of disambiguation undemanding on these POS tags. Since BERT + Dense attained overall better results than ELMo and BERT + Bi-LSTM, in what follows we report its performance only. Finally, to better analyse the impact that the two semantically-empty labels of CSI and WordNet Domains have on the results, we compared the precision, recall and F1 obtained when excluding GENERAL (CSI) and FACTOTUM (WND) from the valid answers. As shown in Table 7, CSI obtains a higher precision and recall, thus confirming our hunch that FACTOTUM highly impacts WordNet Domains performance. In fact, most testing instances were tagged with FACTOTUM by the model and when it was excluded from the valid answers, this made the recall drop.

We note that, when taking full advantage of the CSI labels and let BERT + Dense train on all SemCor instances covered by CSI, we report an 85.9 F1 score on ALL. This, together with the results of the qualitative analysis (Section 4.2), highlights that CSI is the most viable candidate to replace or complement fine-grained inventories for WSD.

4.4 Zero- and Few-Shot Learning

To investigate the improvement that CSI can bring to the disambiguation of unseen or under-represented words, we performed a zero- and few-shot learning experiment. We randomly sampled a set of words, and trained the WSD model removing the annotations for those words. Then, we tested the ability of the class-based sense inventories to leverage labels from other words when zero or only few annotated examples for a word are provided.

Experimental Setup We define L as the set of all lemmas appearing in SemCor and the evaluation datasets. From L we sample a set L_{out} of 100 words, that we partition in two disjoint subsets, L_{test} and L_{dev}, of size 70 and 30 respectively. We define D^W as the subset of dataset D which contains only instances for the lemmas in a set W. For example, SemCor^L is the unmodified training corpus, and Senseval-$2^{L_{dev}}$ is the dataset containing all the instances in Senseval-2 for lemmas in L_{dev}. Starting from SemCor, we build the training set SemCor$^{L \setminus L_{out}}$, i.e., containing as instances all lemmas not in L_{out}, which we used for training the BERT + Dense WSD model with the hyperparameters defined in Section 4.3; we will refer to this model as M_0. To perform

Inventory	F1			PPL			GTO			MFS
	T_0	T_3	T_5	T_0	T_3	T_5	T_0	T_3	T_5	
CSI	$69.0 \pm 2 \times 10^{-4}$	$68.6 \pm 8 \times 10^{-5}$	$77.8 \pm 7 \times 10^{-4}$	4.88	1.54	1.85	**1.84**	**1.03**	**1.20**	72.1
WND	$64.3 \pm 1 \times 10^{-3}$	$75.1 \pm 2 \times 10^{-4}$	$76.2 \pm 3 \times 10^{-5}$	4.41	1.39	1.57	1.68	1.02	1.09	74.9
SuS	$62.6 \pm 3 \times 10^{-4}$	$67.8 \pm 1 \times 10^{-3}$	$73.0 \pm 2 \times 10^{-3}$	4.07	1.51	1.73	1.60	1.01	1.12	68.7

Table 6: Comparison of CSI against WordNet Domains (WND) and SuperSenses (SuS) on zero- and few-shot settings.

CSI vs. WordNetDomains						
Dataset	Precision		Recall		F1	
	CSI	WND	CSI	WND	CSI	WND
Senseval-2	96.2	95.5	86.8	85.3	**91.2**	90.1
Senseval-3	97.0	91.2	85.0	78.6	**90.6**	84.4
SemEval-07	96.8	87.2	73.4	65.4	**83.5**	74.7
SemEval-13	96.7	93.9	84.0	79.0	**89.9**	85.8
SemEval-15	96.8	96.2	75.8	66.7	**85.0**	78.8
ALL	96.6	94.0	84.0	79.1	**89.9**	85.9

Table 7: Comparison of CSI against WordNet Domains (WND) when discarding semantically-empty predictions.

the tuning and evaluation of the models described below, we use Senseval-$2^{L_{dev}}$ as dev set and ALL$^{L_{test}}$ as test set.

Zero- and Few-Shot Setting We evaluated the performance of M_0 on ALL$^{L_{test}}$ without any further training. Note that the lemmas in L_{test} had never been seen tagged during training, so we called this the zero-shot setting. To evaluate the inventory and the model on the few-shot learning task, instead, we built the training datasets T_3, T_5 by randomly sampling $3, 5$ examples, respectively, for each lemma in L_{test}, such that $T_3 \subset T_5$. For each T_i we trained a separate model M_i that we tuned on Senseval-$2^{L_{dev}}$ as done for M_0. Finally, we initialized the weights from M_0 and back-propagated the gradients only through the dense layer.

Evaluation Measure The experiment aims at proving that, even if the training set contains only a few tagged examples for a word, the model can still benefit from the class-based nature of CSI in classifying the under-represented words. Therefore, we measure the performance of the models in terms of both F1 and GTO. Each inventory is compared against its most frequent sense (MFS) baseline, which, given a target word w, is defined as the class that is most frequently used to tag w in SemCor. Since we sample the test lemmas, L_{test}, at random, we report the average of the results obtained on three random word samples L^1_{test}, L^2_{test} and L^3_{test} on their respective instances in the ALL dataset, i.e., ALL$^{L^1_{test}}$, ALL$^{L^2_{test}}$ and ALL$^{L^3_{test}}$.

Results In Table 6 we compare CSI, WordNet Domains and SuperSenses with their MFS baselines. While all the sense inventories manage to beat their MFS, CSI is the one that surpasses it with the greatest gap, i.e., 5.7 F1 points compared to the 1.3 and 4.3 for WordNet Domains and SuperSenses, respectively. This proves that CSI allows the network to better exploit the semantic information carried by the words within the training set, hence enabling a model to generalise well over under-represented words and mitigating the need for large amounts of annotated data for WSD. This is further confirmed when considering the GTO scores in Table 6, where CSI reaches the best performance across the board. Therefore, not only does CSI lead the WSD model to attain higher results than its competitors as regards the MFS, but it also provides - in this setting as well - a better balance between polysemy and performances of the model.

5 Conclusion

In this paper we presented CSI, a new sense inventory for coarse-grained WSD. Our labels proved to be of higher quality than those of alternative inventories, as they exhibited a descriptiveness that was not matched by any of the other inventories, hence making the text annotated with CSI labels of easier interpretation for humans. Moreover, we showed that CSI enabled annotators to attain a higher agreement compared to other fine- and coarse-grained inventories that are employed for the task. On the quantitative side, we showed that CSI allows a supervised WSD model to achieve the most competitive trade-off between performance and expressiveness, and to attain almost 86 F1 points overall when not restricting the set of training instances to those also covered by other inventories as well. In addition, we showed that, when using CSI labels, a supervised model can better generalise over rare words, i.e., those that never or seldom appear in the training data. In fact, in the few-shot learning task, our inventory was the one that led the underlying model to achieve the highest increment over the MFS when just five training examples were provided for the tested words. Foreseeing the potential benefits that CSI can bring to coarse-grained WSD, we release to the community the full inventory, covering more than 120K unique words in the English vocabulary, together with its mapping to WordNet synsets and the code to reproduce the experiments at http://lcl.uniroma1.it/csi.

Acknowledgments

The authors gratefully acknowledge the support of the ERC Consolidator Grant MOUSSE No. 726487 under the European Union's Horizon 2020 research and innovation programme.

This work was supported in part by the MIUR under grant "Dipartimenti di eccellenza 2018-2022" of the Department of Computer Science of the Sapienza University of Rome.

References

Bevilacqua, M., and Navigli, R. 2019. Quasi Bidirectional Encoder Representations from Transformers for Word Sense Disambiguation. In *Proc. of RANLP*, 122–131.

Camacho-Collados, J., and Navigli, R. 2017. BabelDomains: Large-Scale Domain Labeling of Lexical Resources. In *Proc. of EACL*, 223–228.

Cohen, J. 1960. A Coefficient of Agreement of Nominal Scales. *Educational and Psychological Measurement* 20(1):37–46.

Delli Bovi, C.; Espinosa Anke, L.; and Navigli, R. 2015. Knowledge Base Unification via Sense Embeddings and Disambiguation. In *Proc. of EMNLP*, 726–736.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proc. of NAACL*, 4171–4186.

Dewey, M. 1876. *A Classification and Subject index, for Cataloguing and Arranging the Books and Pamphlets of a Library*. Brick Row Book Shop, Incorporated.

Di Eugenio, B., and Glass, M. 2004. The Kappa Statistic: A Second Look. *Computational Linguistics* 30(1):95–101.

Edmonds, P., and Kilgarriff, A. 2002. Introduction to the Special Issue on Evaluating Word Sense Disambiguation Systems. *Natural Language Engineering* 8(4):279–291.

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

Hovy, E.; Marcus, M.; Palmer, M.; Ramshaw, L.; and Weischedel, R. 2006. OntoNotes: The 90% Solution. In *Proc. of NAACL*, 57–60.

Izquierdo, R.; Suárez, A.; and Rigau, G. 2007. Exploring the Automatic Selection of Basic Level Concepts. In *Proc. of RANLP*.

Izquierdo, R.; Suarez, A.; and Rigau, G. 2015. Word vs. Class-Based Word Sense Disambiguation. *Journal of Artificial Intelligence Research* 54:83–122.

Kilgarriff, A. 1997. What is Word Sense Disambiguation Good For? In *Proc. of NLPRS*, 209–214.

Kingma, D. P., and Ba, J. 2015. Adam: A Method for Stochastic Optimization. In *Proc. of ICLR*.

Kipfer, B. A., and Chapman, R. L. 2011. *Roget's International Thesaurus*. Collins Reference.

Komninos, A., and Manandhar, S. 2016. Structured Generative Models of Continuous Features for Word Sense Induction. In *Proc. of COLING*, 3577–3587.

Kraemer, H. C. 1980. Extension of the Kappa Coefficient. *Biometrics* 36(2):207–216.

Landis, J. R., and Koch, G. G. 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics* 159–174.

Magnini, B., and Cavaglià, G. 2000. Integrating Subject Field Codes into WordNet. In *Proc. of LREC*, 1413–1418.

Miller, G. A.; Chodorow, M.; Landes, S.; Leacock, C.; and Thomas, R. G. 1993. Using a Semantic Concordance for Sense Identification. In *Proc. of the Workshop on Human Language Technology*, 240–243.

Moro, A., and Navigli, R. 2015. SemEval-2015 Task 13: Multilingual All-Words Sense Disambiguation and Entity Linking. In *Proc. of SemEval*, 288–297.

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.; Jurgens, D.; and Vannella, D. 2013. SemEval-2013 Task 12: Multilingual Word Sense Disambiguation. In *Proc. of SemEval*, 222–231.

Navigli, R.; Litkowski, K. C.; and Hargraves, O. 2007. SemEval-2007 Task 07: Coarse-grained English All-words Task. In *Proc. of SemEval*, 30–35.

Navigli, R. 2006. Meaningful Clustering of Senses Helps Boost Word Sense Disambiguation Performance. In *Proc. of ACL*, 105–112.

Navigli, R. 2009. Word Sense Disambiguation: A survey. *ACM Computing Surveys* 41(2):1–69.

Navigli, R. 2018. Natural Language Understanding: Instructions for (Present and Future) Use. In *Proc. of IJCAI*, 5697–5702.

Palmer, M.; Babko-Malaya, O.; and Dang, H. T. 2004. Different Sense Granularities for Different Applications. In *Proc. of the NAACL-HLT Workshop on Scalable Natural Language Understanding Systems*, 49–56.

Palmer, M.; Fellbaum, C.; Cotton, S.; Delfs, L.; and Dang, H. T. 2001. English Tasks: All-Words and Verb Lexical Sample. In *Proc. of SENSEVAL-2*, 21–24.

Palmer, M.; Dang, H. T.; and Fellbaum, C. 2007. Making Fine-Grained and Coarse-Grained Sense Distinctions, both Manually and Automatically. *Natural Language Engineering* 13(2):137–163.

Pasini, T., and Navigli, R. 2018. Two knowledge-based methods for high-performance sense distribution learning. In *Proceedings of AAAI*, 5374–5381.

Peters, M. E.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; and Zettlemoyer, L. 2018. Deep Contextualized Word Representations. In *Proc. of NAACL-HLT*, 2227–2237.

Pradhan, S. S.; Loper, E.; Dligach, D.; and Palmer, M. 2007. SemEval-2007 Task 17: English Lexical Sample, SRL and All Words. In *Proc. of SemEval-2007*, 87–92.

Pu, X.; Pappas, N.; Henderson, J.; and Popescu-Belis, A. 2018. Integrating Weakly Supervised Word Sense Disambiguation into Neural Machine Translation. *TACL* 6:635–649.

Raganato, A.; Camacho-Collados, J.; and Navigli, R. 2017. Word Sense Disambiguation: A Unified Evaluation Framework and Empirical Comparison. In *Proc. of EACL*, 99–110.

Raganato, A.; Delli Bovi, C.; and Navigli, R. 2017. Neural Sequence Learning Models for Word Sense Disambiguation. In *Proc. of EMNLP*, 1167–1178.

Scarlini, B.; Pasini, T.; and Navigli, R. 2019. Just OneSec for Producing Multilingual Sense-Annotated Data. In *Proc. of ACL*, 699–709.

Scarlini, B.; Pasini, T.; and Navigli, R. 2020. SensEmBERT: Context-Enhanced Sense Embeddings for Multilingual Word Sense Disambiguation. In *Proc. of AAAI*.

Snow, R.; Prakash, S.; Jurafsky, D.; and Ng, A. Y. 2007. Learning to Merge Word Senses. In *Proc. of ACL*, 1005–1014.

Snyder, B., and Palmer, M. 2004. The English All-Words Task. In *Proc. of Senseval-3*, 41–43.

Vial, L.; Lecouteux, B.; and Schwab, D. 2019. Sense Vocabulary Compression through the Semantic Knowledge of WordNet for Neural Word Sense Disambiguation. In *Proc. of GWC*.