Sense-Annotated Corpora for Word Sense Disambiguation in Multiple Languages and Domains

Bianca Scarlini, Tommaso Pasini, Roberto Navigli
Sapienza NLP Group
Department of Computer Science
Sapienza University of Rome
{scarlini, pasini, navigli}@di.uniroma1.it

1. Introduction

Nowadays, deep neural models use data as their “fuel”, and have proved to attain better performance the larger the training corpora they are provided with. Unfortunately, producing large manually-curated corpora is an expensive and time-consuming task. This is especially an issue for the field of Natural Language Processing (NLP), where data is usually tied to a specific language, thereby further increasing the time and cost of producing large-scale annotated corpora. An area of NLP that particularly suffers from such a problem is Word Sense Disambiguation (WSD), that is, the task of assigning a meaning to a word in a given context (Navigli, 2009). It is not by chance, in fact, that corpora annotated with senses are limited for English and almost absent for other languages. Indeed, while various sense-annotated corpora do exist for English, most of these are limited in terms of words and senses covered: MASC (Passonneau et al., 2012), a corpus providing 1M instances manually-annotated with senses, covers only 100 distinct words, OntoNotes (Hovy et al., 2006), instead, provides annotated examples for only 6500 concepts of less than 3500 words.

SemCor (Miller et al., 1993) is the largest manually-annotated English corpus currently available. It contains roughly 40K sentences and 200K content words annotated with their corresponding meanings. Despite being the standard training corpus for WSD, SemCor’s main limitation resides in its low coverage of the English vocabulary in terms of both words and meanings. In fact, it provides annotations for only 22K unique lexemes, i.e., lemma and POS tag pairs, which cover not even 15% of the words comprised in WordNet (Miller et al., 1990), i.e., the largest and most-used electronic dictionary for English. Moreover, as we shift our focus towards lower-resourced languages, the need for semantically-annotated data becomes increasingly urgent. Indeed, SemCor, with some exceptions (Bentivogli and Pianta, 2005; Bond et al., 2012), lacks an adequate multilingual counterpart in most world languages and, hence, WSD models are limited when it comes to scaling over languages other than English. In this scenario, several automatic approaches for producing multilingual sense-annotated data (Pasini and Navigli, 2017; Pasini et al., 2018; Scarlini et al., 2019; Pasini and Navigli, 2020) have tried to mitigate the aforementioned shortcomings. In fact, being able to create silver annotated data on a large scale is crucial to freeing WSD models from dependence on exclusively those words and languages that are comprised within manually-curated resources.

In this paper we follow Scarlini et al. (2019) and apply their approach to all the nominal lexemes of 5 major European languages, i.e., English, Italian, Spanish, French and German, so as to ensure full coverage in terms of both words and senses. We release to the community training corpora in 5 different languages, including more than 15 million annotations. Our automatically-produced data lead a supervised model to state-of-the-art results in all multilingual WSD tasks, while remaining competitive with manually-curated resources on English. Furthermore, we exploited automatic approaches for inducing the distribution of word senses (Pasini and Navigli, 2018) to build 5 additional datasets for English, each peculiar to a different semantic domain. These domain-specific datasets proved to lead a supervised model to a significant gain in performance when it comes to in-domain WSD evaluation.

2. Related Work

Tackling the Word Sense Disambiguation problem from a supervised point of view has been the focus of several recent works (Luo et al., 2018; Kumar et al., 2019; Bevilacqua and Navigli, 2019). While on the one hand these models proved to attain state-of-the-art results on the English WSD task, on the other hand they still depended heavily on sense-annotated training corpora, and hence were limited to a restricted set of words, senses and languages. In fact,
the line of research focused on new methodologies for automatically creating high-quality sense-annotated datasets is long-standing (Taghipour and Ng, 2015; Raganato et al., 2016; Delli Bovi et al., 2017; Pasini andNavigli, 2017; Pasini et al., 2018; Scarlini et al., 2019; Pasini andNavigli, 2020). These latter five approaches enabled supervised systems to step outside the boundaries of standard English WSD tasks and allowed them to scale on languages where manually-curated resources are not available. Train-O-Matic (Pasini andNavigli, 2017) is a language-independent method for the creation of large-scale senseannotated corpora. By exploiting the information enclosed within a semantic network it is able to provide high-quality annotations of raw sentences, leading a supervised model to competitive results on both English and multilingual WSD tasks. More recently, Scarlini et al. (2019) dropped the requirement to draw on the structure of a semantic network by exploiting a sparse vector representation of senses and Wikipedia’s inner organization into pages and categories. Their automatically-created corpora proved to outperform all other automatic alternatives on English and achieve state-of-the-art results on all multilingual WSD datasets. In this paper we complement the work by Scarlini et al. (2019) (ONESEC) and apply their approach to all the words in the vocabulary of 5 different European languages. Moreover, we couple their approach with an automatic method for inducing word-sense distribution and create 5 different domain-specific datasets for English in order to enable domain-specific WSD on a large scale.

3. Preliminaries

ONESEC relies on three resources for producing sense-tagged corpora, namely, Wikipedia, the largest freely available electronic encyclopedia, BabelNet (Navigli and Ponzetto, 2012), a multilingual semantic network, and NASARI (Camacho-Collados et al., 2016), a sparse representation of BabelNet concepts.

Wikipedia Wikipedia is the largest encyclopedic corpus currently available. It is made up of approximately 300 separate editions, each comprising texts in a specific language. Wikipedia articles describe either abstract concepts or named entities, and are, in their turn, grouped together into Wikipedia categories. These categories collect a set of articles that share a common feature or characteristic, and organise the information in Wikipedia into macro-areas. For example, the category MODERNIST WRITERS contains, among others, the pages Ernest Hemingway and F. Scott Fitzgerald. For ease of reading, we define a lemma occurring in a category as the lemma that appears in a sentence of any page within that category and the sentences of a category as the sentences of all the pages belonging to the reference category.

BabelNet BabelNet is a multilingual semantic network built by merging together several heterogeneous resources, such as WordNet, Wikipedia, Wikidata etc. It is organised in the form of a graph, where nodes are concepts and edges are semantic relations between them. Each concept combines lexicalizations in different languages. For example, the season concept of spring#n also includes the terms printemps (French), Frühling (German) and primavera (Italian).

NASARI NASARI vectors are sparse-vector representations of BabelNet concepts. Each dimension corresponds to a word that it is associated with a lexical specificity value (Lafon, 1980), which expresses how much the target word is relevant for the concept it represents. For example, the terms calendar, year and season are included in the NASARI vector corresponding to the season sense of spring#n. NASARI proved to be effective in different tasks, such as text classification (Sinoara et al., 2019) and Word Sense Disambiguation (Camacho-Collados et al., 2016). We note that the NASARI lexical vectors are available for nominal concepts only.

4. ONESEC

We now move on to describe ONESEC, which, by relying on the previously introduced resources, automatically produces sense-annotated data in multiple languages. It extends the “One Sense per Discourse” (Gale et al., 1992) assumption to “One Sense per Wikipedia Category”, i.e., all the occurrences of a word in the same category share a common meaning, and leverages the information in a Wikipedia category to automatically tag the occurrences of the words therein. Given a target lexeme l, ONESEC produces annotated examples for each of its senses by applying the following three steps.

Category Representation. First, we collect all the Wikipedia categories C1 . . . Cn such that l appears at least t times across the sentences of their pages. Then, for each category C of the lemma l, we create a Bag of Words representation BlC, i.e., a sparse vector in which dimensions are words scored by their frequency in the sentences of C where l appears.

Sense Assignment. We assign a sense to each lexeme-category pair (l, C) by leveraging the NASARI lexical vectors (see Section 3.). That is, as first step we compute the similarity between BlC and each NASARI vector vl associated with a sense s of l by means of the Weighted Overlap measure. The WO of two vectors v1 and v2 is computed as follows:

$$WO(v_1, v_2) = \ln(|I|+1) \left( \sum_{w \in l} \frac{1}{r_w^{v_1} + r_w^{v_2}} \left( \sum_{i=1}^{|I|} \frac{1}{2i} \right)^{-1} \right)$$

where r_w^{vl} is the rank of the word w in the vector vl and I is the set of intersecting dimensions of v1 and v2. Then, we assign to (l, C) the sense that maximizes the WO similarity with BlC. Formally, being v_s1 . . . v_sn the NASARI vectors of the senses of l and

\[\text{https://www.wikipedia.org/}\]
\[\text{https://babelnet.org/}\]
\[\text{http://nlp.uniroma1.it/nasari/}\]

\[4\]We use the notation lemma#pos.

\[5\]We preferred the Weighted Overlap over the most common Cosine Similarity as it has been proved to be more effective in capturing the distance between sparse vectors (Pilehvar et al., 2013).
Table 1: Statistics on the English corpora restricted to the nominal words of SemCor, SemCor+OMSTI, Train-O-Matic and ONESEC in terms of number of sentences, annotations, unique nouns and unique nominal senses.

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Sentences</th>
<th>Annotations</th>
<th>Distinct Nouns</th>
<th>Distinct Senses</th>
</tr>
</thead>
<tbody>
<tr>
<td>SemCor</td>
<td>18,531</td>
<td>87,002</td>
<td>11,391</td>
<td>15,906</td>
</tr>
<tr>
<td>SemCor+OMSTI</td>
<td>455,762</td>
<td>558,383</td>
<td>11,396</td>
<td>16,194</td>
</tr>
<tr>
<td>Train-O-Matic</td>
<td>12,722,530</td>
<td>12,722,530</td>
<td>51,395</td>
<td>56,229</td>
</tr>
<tr>
<td>ONESEC</td>
<td>8,813,642</td>
<td>8,813,642</td>
<td>28,383</td>
<td>39,848</td>
</tr>
</tbody>
</table>

\[ WO(B_{C_l}^{C_t}, v_{s_i}) = \max_{s_i} \text{arg}(WO(B_{C_l}^{C_t}, s_i)) \]

At the end of this step, each pair \((l, C)\) is mapped to the potentially most suitable sense \(s\) of \(l\).

**Sentence Sampling.** Given the mapping between senses and categories we computed in the previous step, we now select the set of training examples for each sense \(s_i\) of the target lexeme \(l\). To this end, we first set the number of sentences to draw for \(s_i\) by applying a Zipfian distribution as follows:

\[
\mathcal{K}_{s_i} = \mathcal{K}(i^z)^{-1}
\]

where \(i\) is the rank of \(s_i\) according to BabelNet’s ordering of senses, \(\mathcal{K}\) is a parameter of the system and determines the number of examples for the first sense of \(l\) and \(z\) is the value of the exponent that controls how fast the distribution decreases. Then, for each sense \(s_i\) of \(l\) we take all the categories in which \(l\) appears that are associated with \(s_i\) and sample a number of sentences for each category that is proportional to its weighted overlap similarity with respect to \(s_i\). Formally, given the list of categories \(C_1 \ldots C_m\) associated with \(s_i\) and ordered according to their weighted overlap score, we define the number of sentences to draw from \(C_j\) as follows:

\[
\mathcal{K}_{s_iC_j} = \mathcal{K}_{s_i} \frac{1}{j} \left( \sum_{j=1}^{m} \frac{1}{j} \right)^{-1}
\]

Finally, for each category \(C_j\) we perform a weighted sampling in which the probability of a sentence being selected is determined by its perplexity, i.e., the lower the perplexity the higher the probability of a sentence being sampled. At the end of this procedure every occurrence of \(l\) in the selected sentences is tagged with the sense \(s_i\).

**5. Creating the Corpora**

We apply ONESEC’s procedure to all the nominal words in the vocabulary of five major European languages, i.e., English, Italian, Spanish, French and German. More in detail, as regards English, we consider the whole set of nouns of WordNet. As for the other languages, we take into account the set of nouns in the WordNet part of BabelNet, i.e., we take all the synsets in BabelNet that contain a WordNet sense and collect all the nouns in the target language therein.

Moreover, we slightly modify ONESEC to create domain-specific datasets for all English nominal lexemes by exploiting DaD (Pasini et al., 2018), i.e., a knowledge-based method for inducing in-domain sense distributions. DaD computes the distribution over BabelNet synsets by first approximating the probability of a BabelDomain (Camacho-Collados and Navigli, 2017) being a topic covered in the corpus of raw sentences, and then leverages the connections in BabelNet to propagate the domains’ probabilities over all BabelNet synsets. Hence, in order to select the number of sentences to assign to each sense of a lemma \(l\) (see Equation 1), for each domain \(d\) we rank the senses of \(l\) according to DaD’s sense distributions associated with \(d\), instead of following BabelNet’s ordering of senses. We create sense-annotated corpora according to the domain-aware sense distributions for 5 different domains, i.e., Biology, Health Care, Politics, Social Issues and Maths & PC.

We note that for building ONESEC’s corpora we have to set 3 different system parameters: i) \(t\), i.e., the minimum number of occurrences of a lemma in a Wikipedia category, ii) \(K\), i.e., the maximum number of sentences to be assigned with the first sense of a lemma, iii) \(z\), i.e., the exponent that controls the Zipf’s distribution (see Section 4). For both English and multilingual datasets, we set \(t = 20\). In the case that we find no categories for a lemma with \(t = 20\), we relax the constraint and set \(t = 10\). As for the other parameters, we follow Scarlini et al. (2019) and set \(K = 700\) and \(z = 2.1\) for English and \(K = 200\) and \(z = 2.0\) for the other languages.

**6. Statistics**

In this Section we show the statistics of our corpora in order to give a general overview of them. In Table 2 we report the characteristics of each corpus divided per language. As can be seen, the English corpus is the richest. The average polysemy of ONESEC on English, in fact, is higher than the one of WordNet\(^6\) meaning that the lemmas covered by ONESEC are among those with the highest polysemy and most of the lemmas that are not covered are, instead, monosemous. The average number of sentences per sense is higher for English than for the other languages inasmuch as it directly depends on the parameter \(K\) (Section 5). The French corpus is the second richest one in terms of number of annotations and lemmas and senses covered. In to-

---

\(^6\)The perplexity is computed by means of the Neural Language Model presented in Howard and Ruder (2018).
As can be seen in Table 1 when compared to other manually- and semi-automatically annotated corpora, i.e., SemCor (Miller et al., 1993) and SemCor+OMSTI (Taghipour and Ng, 2015), ONESEC covers more than double their lemmas and senses with one order of magnitude more annotations. Compared to SemCor+OMSTI, in fact, ONESEC covers almost three times more nouns and two times more senses. Furthermore, OMSTI covers roughly the same number of nouns as SemCor with a slightly higher number of meanings. Train-O-Matic, instead, is the training corpus with the largest coverage in terms of both lemmas and senses. One reason why ONESEC falls behind Train-O-Matic in terms of coverage is that it filters out all the categories that contain less than 10 sentences in which the target word occurs (see Section 4 and 5). Therefore, it may happen that a word-sense pair has no category associated with it and hence no annotations can be provided.

Even though ONESEC covers less nouns, it is more accurate, as shown in our experiments. ONESEC, in fact, leads a supervised WSD model to attain higher results in comparison to the same WSD model trained on Train-O-Matic data. In Table 4 we break down the statistics for each corpus we created for 5 distinct semantic domains on English. As can be seen, ONESEC can find as many examples in each domain as in the general-domain corpus (Table 2), meaning that it can also cover senses that are, instead, peculiar to a specific domain.

### 7. Experimental Setup

For assessing the quality of ONESEC annotations we trained a supervised WSD model on our automatically-generated data and tested its performance on 5 standard WSD benchmarks for English and for another four languages, i.e., Italian, Spanish, French and German.

#### Reference WSD models

As for English, we used It Makes Sense (Zhong and Ng, 2010) IMS, a support vector machine-system which builds a single model for each target word in the training set. For the other languages, instead, we employed the BiLSTM-based model proposed by Raganato et al. (2017b).

#### Test Data

We tested on all the nominal instances comprised in the 5 standard English datasets included in the framework of Raganato et al. (2017a), namely, Senseval-2 (Edmonds and Cotton, 2001), Senseval-3 (Snyder and Palmer, 2004), SemEval-07 (Pradhan et al., 2007), SemEval-13 (Navigli et al., 2013) and SemEval-15 (Moro and Navigli, 2015). Furthermore, we report the results on the ALL dataset, i.e., the concatenation of all the aforementioned test sets. For evaluating the corpora in the other four languages, instead, we tested the reference WSD model (BiLSTM) on the multilingual datasets of SemEval-2013 task 12 (Navigli et al., 2015) (Italian, Spanish, French and German) and SemEval-2015 task 13 (Moro and Navigli, 2015) (Italian and Spanish).

#### Domain-Specific Evaluation

To evaluate the domain-specific corpora produced by ONESEC we generated a specific training corpus (ONESEC$_{dom}$) for each of the following five domains: Biology, Health Care, Politics, Social Issues and Maths & PC, as explained at the end of Section 5. Then, we trained IMS on each training set separately, and tested it on the documents of SemEval-13 and SemEval-15 corresponding to each of the aforementioned domains.
<table>
<thead>
<tr>
<th>Dataset</th>
<th>Domain</th>
<th>Size</th>
<th>Backoff</th>
<th>(\text{ONESeC}_{\text{dom}})</th>
<th>(\text{ONESeC})</th>
<th>TOM</th>
<th>MFS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>P</td>
<td>R</td>
<td>F1</td>
<td>P</td>
</tr>
<tr>
<td>SemEval-13</td>
<td>Biology</td>
<td>135</td>
<td>MFS</td>
<td>80.7</td>
<td>80.7</td>
<td>80.7</td>
<td>67.4</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>-</td>
<td>79.4</td>
<td>74.1</td>
<td>76.6</td>
<td>65.1</td>
</tr>
<tr>
<td></td>
<td>Health Care</td>
<td>138</td>
<td>MFS</td>
<td>67.4</td>
<td>67.4</td>
<td>67.4</td>
<td>65.9</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>-</td>
<td>64</td>
<td>50.8</td>
<td>60.8</td>
<td>62.4</td>
</tr>
<tr>
<td></td>
<td>Politics</td>
<td>279</td>
<td>MFS</td>
<td>73.5</td>
<td>73.5</td>
<td>73.5</td>
<td>68.8</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>-</td>
<td>72.0</td>
<td>68.1</td>
<td>70.0</td>
<td>67.0</td>
</tr>
<tr>
<td></td>
<td>Social Issues</td>
<td>349</td>
<td>MFS</td>
<td>73.6</td>
<td>73.6</td>
<td>73.6</td>
<td>66.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>-</td>
<td>70.7</td>
<td>62.2</td>
<td>66.2</td>
<td>62.5</td>
</tr>
<tr>
<td>SemEval-15</td>
<td>Maths &amp; Pc</td>
<td>97</td>
<td>MFS</td>
<td>63.0</td>
<td>63.0</td>
<td>63.0</td>
<td>60.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>-</td>
<td>62.9</td>
<td>63.0</td>
<td>61.9</td>
<td>59.8</td>
</tr>
</tbody>
</table>

Table 5: Domain-specific evaluation on SemEval-2013 and SemEval-2015 of IMS trained on \(\text{ONESeC}_{\text{dom}}\), \(\text{ONESeC}\) and TOM.

<table>
<thead>
<tr>
<th>Lang</th>
<th>(\text{ONESeC})</th>
<th>TOM - Bi-LSTM</th>
<th>Best</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P</td>
<td>R</td>
<td>F1</td>
</tr>
<tr>
<td>IT</td>
<td>72.3</td>
<td>64.5</td>
<td>68.2</td>
</tr>
<tr>
<td>ES</td>
<td>76.0</td>
<td>68.3</td>
<td>72.0</td>
</tr>
<tr>
<td>FR</td>
<td>79.2</td>
<td>70.9</td>
<td>74.8</td>
</tr>
<tr>
<td>DE</td>
<td>83.0</td>
<td>68.5</td>
<td>75.1</td>
</tr>
</tbody>
</table>

Table 6: Comparison of Bi-LSTM trained on \(\text{ONESeC}\) and TOM with the best system (Best) on SemEval-2013. * Train-O-Matic, * UMCC-DLSI.

<table>
<thead>
<tr>
<th>Lang</th>
<th>(\text{ONESeC})</th>
<th>TOM - Bi-LSTM</th>
<th>Best</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P</td>
<td>R</td>
<td>F1</td>
</tr>
<tr>
<td>IT</td>
<td>65.0</td>
<td>60.2</td>
<td>62.5</td>
</tr>
<tr>
<td>ES</td>
<td>67.8</td>
<td>58.4</td>
<td>62.8</td>
</tr>
</tbody>
</table>

Table 7: Comparison of Bi-LSTM trained on \(\text{ONESeC}\) and TOM with the best system (Best) on SemEval-2015. * Train-O-Matic.

Competitors We compared the results attained by the reference WSD models trained on \(\text{ONESeC}\) and other three training corpora for English:

- **Train-O-Matic** (TOM), a knowledge-based approach for producing training corpora for WSD in English and all the languages supported by BabelNet.
- **OMSTI**, a semi-automatic approach for generating sense-annotated data that leverages parallel corpora and manual annotations.
- **SemCor**, a manually-annotated corpus which is the de-facto standard training set for English Word Sense Disambiguation models.

As for the multilingual setting, we compared \(\text{ONESeC}\) with the best performing model in each of the tested datasets and with the reference multilingual WSD model (Bi-LSTM) when trained on Train-O-Matic corpora.

8. Results

As a first result in Table 6 we report the performance of IMS on the 5 general-domain English WSD benchmarks. As can be seen, \(\text{ONESeC}\) corpora lead IMS to attain the highest results across the board when compared to its automatic and semi-automatic competitors, i.e., Train-O-Matic and OMSTI. \(\text{ONESeC}\), in fact, ranks second only in comparison to SemCor, which, however, is a manually-curated resource. It performs only 1.4 F1 points lower than SemCor on ALL, a loss that is due to the automatic nature of \(\text{ONESeC}\) and to the unavoidable noise that can be found in silver data. Nevertheless, we note that \(\text{ONESeC}\) outperforms SemCor in 2 out of 5 of the tested datasets, i.e., SemEval-13 and SemEval-15, with an increment of 1 and 4.7 F1 points, respectively.

In Table 7 we report the results on the domains of SemEval-13 and SemEval-15 attained by IMS trained on the 5 \(\text{ONESeC}_{\text{dom}}\) corpora and compare them with the results achieved when training IMS on the general-domain training corpora of \(\text{ONESeC}\) and Train-O-Matic (TOM). \(\text{ONESeC}_{\text{dom}}\) leads IMS to attain the best results across each domain, with the highest boost of more than 13 points compared to its general version on the Biology domain, and of more than 17 points compared to Train-O-Matic. This shows the effectiveness of \(\text{ONESeC}\) in providing examples for senses that are specific to a given domain.

As regards the multilingual evaluation, in Tables 6 and 7 we report the results attained by the BiLSTM-based model on each specific language, separately. To set a level playing field with Train-O-Matic, which reported the results attained by IMS on the multilingual evaluation, we trained the same BiLSTM-based model on their data as well. \(\text{ONESeC}\) attains, also in this case, the highest results across the board, repeatedly beating both Train-O-Matic and the best performing system (UMCC-DLSI (Gutiérrez et al., 2010)) on each language by several points.

Overall, our corpora proved to be of high quality in different settings and languages. In fact, \(\text{ONESeC}\) places itself as the best alternative when it comes to train a model on a specific domain or on lower-resourced languages, while remaining competitive with manually-curated resources on English.

9. Conclusion

In this paper we propose a total of 10 different datasets automatically annotated with sense labels for 5 different languages and 5 distinct semantic domains. Our experi-
ments proved that our datasets are of high quality and can be used as training data by supervised models to perform Word Sense Disambiguation in each of the 5 different languages. Furthermore, when it comes to in-domain WSD, our domain-specific datasets proved to be effective in all the tested scenarios, aiding WSD approaches to perform better in each specific domain.

As future work, we plan to refine our approach by exploiting other knowledge resources, such as SensEmBERT (Scarlini et al., 2020), VerbAtlas (Di Fabio et al., 2019) or SyntagNet (Maru et al., 2019).

We release all the data at http://trainomatic.org/onesec comprising more than 15 million annotations across 5 different languages. We plan to include more languages and domains in the future.

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.

10. Bibliographical References


