4	Very similar	The two words are synonyms (e.g., *midday-noon* or *motherboard-mainboard*).
3	Similar	The two words share many of the important ideas of their meaning but include slightly different details. They refer to similar but not identical concepts (e.g., *lion-zebra* or *firefighter-policeman*).
2	Slightly similar	The two words do not have a very similar meaning, but share a common topic/domain/function and ideas or concepts that are related (e.g., *house-window* or *airplane-pilot*).
1	Dissimilar	The two words describe clearly dissimilar concepts, but may share some small details, a far relationship or a domain in common and might be likely to be found together in a longer document on the same topic (e.g., *software-keyboard* or *driver-suspension*).
0	Totally dissimilar and unrelated	The two words do not mean the same thing and are not on the same topic (e.g., *pencil-frog* or *PlayStation-monarchy*).

Table 2: The five-point Likert scale used to rate the similarity of item pairs. See Table 4 for examples.

Similarity (Jurgens et al., 2014) which was designed to systematically order a broad range of semantic relations: synonymy, similarity, relatedness, topical association, and unrelatedness. Table 2 describes the five points in the similarity scale along with example word pairs.

Pairing word selection. Having the initial 500-word seed set at hand, we selected a pair for each word. The selection was carried out in such a way as to ensure a uniform distribution of pairs across the similarity scale. In order to do this, we first assigned a random intended similarity to each pair. The annotator then had to pick the second word so as to match the intended score. In order to allow the annotator to have a broader range of candidate words, the intended score was considered as a similarity interval, one of [0-1], [1-2], [2-3] and [3,4]. For instance, if the first word was *helicopter* and the presumed similarity was [3-4], the annotator had to pick a pairing word which was "semantically similar" (see Table 2) to *helicopter*, e.g., *plane*. Of the 500 pairing words, 45 (9%) and 71 (14%) were named entities and multiwords, respectively. This resulted in an English dataset comprising 500 word pairs, 105 (21%) and 112 (22%) of which have at least one named entity and multiword, respectively.

2.1.2 Dataset translation

The remaining four multilingual datasets (i.e., Farsi, German, Italian, and Spanish) were constructed by translating words in the English dataset to the target language. We had two goals in mind while selecting translation as the construction strategy of these datasets (as opposed to independent word samplings per language): (1) to have comparable datasets across languages in terms of domain coverage, multiword and named entity distribution[4] and (2) to enable an automatic construction of cross-lingual datasets (see Section 2.2).

Each English word pair was translated by two independent annotators. In the case of disagreement, a third annotator was asked to pick the preferred translation. While translating, the annotators were shown the word pair along with their initial similarity score, which was provided to help them in selecting the correct translation for the intended meanings of the words.

2.1.3 Scoring

The annotators were instructed to follow the guidelines, with special emphasis on distinguishing between similarity and relatedness. Furthermore, although the similarity scale was originally designed as a Likert scale, annotators were given flexibility to assign values between the defined points in the scale (with a step size of 0.25), indicating a blend of two relations. As a result of this procedure, we obtained 500 word pairs for each of the five languages. The pairs in each language were shuffled and their initial scores were discarded. Three annotators were then asked to assign a similarity score to each pair according to our similarity scale (see Section 2.1.1).

Table 3 (first row) reports the average pairwise Pearson correlation among the three annotators for each of the five languages. Given the fact that our word pairs spanned a wide range of domains, and that there was a possibility for annotators to misunderstand some words, we devised a procedure to check the quality of the annotations and to improve the reliability of the similarity scores. To this end, for each dataset and for each annotator

[4] Apart from the German dataset in which the proportion of multiwords significantly reduces (from 22% of English to around 11%) due to the compounding nature of the German language, other datasets maintain similar proportions of multiwords to those of the English dataset.

	English	Farsi	German	Italian	Spanish
Initial scores	0.836	0.839	0.864	0.798	0.829
Revised scores	0.893	0.906	0.916	0.900	0.890

Table 3: Average pairwise Pearson correlation among annotators for the five monolingual datasets.

MONOLINGUAL			
DE	Tuberkulose	LED	0.25
ES	zumo	batido	3.00
EN	Multiple Sclerosis	MS	4.00
IT	Nazioni Unite	Ban Ki-moon	2.25
FA	شام آخر	لئوناردو دا وینچی	2.08
CROSS-LINGUAL			
DE-ES	Sessel	taburete	3.08
DE-FA	Lawine	برف	2.25
DE-IT	Taifun	ciclone	3.46
EN-DE	pancreatic cancer	Chemotherapie	1.75
EN-ES	Jupiter	Mercurio	3.25
EN-FA	film	پوچ‌گرایی	0.25
EN-IT	island	penisola	3.08
ES-FA	duna	بیابان	2.25
ES-IT	estrella	pianeta	2.83
IT-FA	avvocato	نمایشگر	0.08

Table 4: Example pairs and their ratings (EN: English, DE: German, ES: Spanish, IT: Italian, FA: Farsi).

we picked the subset of pairs for which the difference between the assigned similarity score and the average of the other two annotations was more than 1.0, according to our similarity scale. The annotator was then asked to revise this subset performing a more careful investigation of the possible meanings of the word pairs contained therein, and change the score if necessary. This procedure resulted in considerable improvements in the consistency of the scores. The second row in Table 3 ("Revised scores") shows the average pairwise Pearson correlation among the three revised sets of scores for each of the five languages. The inter-annotator agreement for all the datasets is consistently in the 0.9 ballpark, which demonstrates the high quality of our multilingual datasets thanks to careful annotation of word pairs by experts.

2.2 Cross-lingual datasets

The cross-lingual datasets were automatically created on the basis of the translations obtained with the method described in Section 2.1.2 and using the approach of Camacho-Collados et al. (2015).[5] By intersecting two aligned translated pairs across two languages (e.g., *mind-brain* in English and *mente-cerebro* in Spanish), the approach creates two cross-lingual pairs between the two languages (*mind-cerebro* and *brain-mente* in the example). The similarity scores for the constructed cross-lingual pairs are computed as the average of the corresponding language-specific scores in the monolingual datasets. In order to avoid semantic shifts between languages interfering in the process, these pairs are only created if the difference between the corresponding language-specific scores is lower than 1.0. The full details of the algorithm can be found in Camacho-Collados et al. (2015). The approach has been validated by human judges and shown to achieve agreements of around 0.90 with human judges, which is similar to inter-annotator agreements reported in Section 2.1.3. See Table 4 for some sample pairs in all monolingual and cross-lingual datasets. Table 5 shows the final number of pairs for each language pair.

[5] http://lcl.uniroma1.it/similarity-datasets/

	EN	DE	ES	IT	FA
EN	500	914	978	970	952
DE	-	500	956	912	888
ES	-	-	500	967	967
IT	-	-	-	500	916
FA	-	-	-	-	500

Table 5: Number of word pairs in each dataset. The cells in the main diagonal of the table (e.g., EN-EN) correspond the monolingual datasets of subtask 1.

3 Evaluation

We carried out the evaluation on the datasets described in the previous section. The experimental setting is described in Section 3.1 and the results are presented in Section 3.2.

3.1 Experimental setting

3.1.1 Evaluation measures and official scores

Participating systems were evaluated according to standard Pearson and Spearman correlation mea-

sures on all word similarity datasets, with the final official score being calculated as the harmonic mean of Pearson and Spearman correlations (Jurgens et al., 2014). Systems were allowed to participate in either multilingual word similarity, cross-lingual word similarity, or both. Each participating system was allowed to submit a maximum of two runs.

For the multilingual word similarity subtask, some systems were multilingual (applicable to different languages), whereas others were monolingual (only applicable to a single language). While monolingual approaches were evaluated in their respective languages, multilingual and language-independent approaches were additionally given a global ranking provided that they tested their systems on at least four languages. The final score of a system was calculated as the average harmonic mean of Pearson and Spearman correlations of the four languages on which it performed best.

Likewise, the participating systems of the cross-lingual semantic similarity subtask were allowed to provide a score for a single cross-lingual dataset, but must have provided results for at least six cross-lingual word similarity datasets in order to be considered for the final ranking. For each system, the global score was computed as the average harmonic mean of Pearson and Spearman correlation on the six cross-lingual datasets on which it provided the best performance.

3.1.2 Shared training corpus

We encouraged the participants to use a shared text corpus for the training of their systems. The use of the shared corpus was intended to mitigate the influence that the underlying training corpus might have upon the quality of obtained representations, laying a common ground for a fair comparison of the systems.

- **Subtask 1.** The common corpus for subtask 1 was the Wikipedia corpus of the target language. Specifically, systems made use of the Wikipedia dumps released by Al-Rfou et al. (2013).[6]

- **Subtask 2.** The common corpus for subtask 2 was the Europarl parallel corpus[7]. This corpus is available for all languages except Farsi. For pairs involving Farsi, participants were allowed to use the OpenSubtitles2016 parallel corpora[8]. Additionally, we proposed a second type of multilingual corpus to allow the use of different techniques exploiting comparable corpora. To this end, some participants made use of Wikipedia.

3.1.3 Participating systems

This task was targeted at evaluating multilingual and cross-lingual word similarity measurement techniques. However, it was not only limited to this area of research, as other fields such as semantic representation consider word similarity as one of their most direct benchmarks for evaluation. All kinds of semantic representation techniques and semantic similarity systems were encouraged to participate.

In the end we received a wide variety of participants: proposing distributional semantic models learnt directly from raw corpora, using syntactic features, exploiting knowledge from lexical resources, and hybrid approaches combining corpus-based and knowledge-based clues. Due to lack of space we cannot describe all the systems in detail, but we recommend the reader to refer to the system description papers for more information about the individual systems: HCCL (He et al., 2017), Citius (Gamallo, 2017), jmp8 (Melka and Bernard, 2017), l2f (Fialho et al., 2017), QLUT (Meng et al., 2017), RUFINO (Jimenez et al., 2017), MERALI (Mensa et al., 2017), Luminoso (Speer and Lowry-Duda, 2017), hhu (QasemiZadeh and Kallmeyer, 2017), Mahtab (Ranjbar et al., 2017), SEW (Delli Bovi and Raganato, 2017) and Wild_Devs (Rotari et al., 2017), and OoO.

3.1.4 Baseline

As the baseline system we included the results of the concept and entity embeddings of NASARI (Camacho-Collados et al., 2016). These embeddings were obtained by exploiting knowledge from Wikipedia and WordNet coupled with general domain corpus-based Word2Vec embeddings (Mikolov et al., 2013). We performed the evaluation with the 300-dimensional English embedded vectors (version 3.0)[9] and used them for all languages. For the comparison within and

[6] https://sites.google.com/site/rmyeid/projects/polyglot

[7] http://opus.lingfil.uu.se/Europarl.php

[8] http://opus.lingfil.uu.se/OpenSubtitles2016.php

[9] http://lcl.uniroma1.it/nasari/