SemEval-2020 Task 2: Predicting Multilingual and Cross-Lingual (Graded) Lexical Entailment

Goran Glavaš[1], Ivan Vulić[2], Anna Korhonen[2] and Simone Paolo Ponzetto[1]
[1] Data and Web Science Group, University of Mannheim
[2] Language Technologies Lab, DTAL, University of Cambridge
{goran,simone}@informatik.uni-mannheim.de
{iv250,alk23}@cam.ac.uk

Abstract

Lexical entailment (LE) is a fundamental asymmetric lexico-semantic relation, supporting the hierarchies in lexical resources (e.g., WordNet, ConceptNet) and applications like natural language inference and taxonomy induction. Multilingual and cross-lingual NLP applications warrant models for LE detection that go beyond language boundaries. As part of SemEval 2020, we carried out a shared task (Task 2) on multilingual and cross-lingual LE. The shared task spans three dimensions: (1) monolingual LE in multiple languages versus cross-lingual LE, (2) binary versus graded LE, and (3) a set of 6 diverse languages (and 15 corresponding language pairs). We offered two different evaluation tracks: (a) distributional (*Dist*): for unsupervised, fully distributional models that capture LE solely on the basis of unannotated corpora, and (b) *Any*: for externally informed models, allowed to leverage any resources, including lexico-semantic networks (e.g., WordNet or BabelNet). In the *Any* track, we received system runs that push state-of-the-art across all languages and language pairs, for both binary LE detection and graded LE prediction.

1 Introduction

Lexical entailment (LE; *hyponymy-hypernymy* or *is-a* relation) is a core asymmetric lexico-semantic relation (Collins and Quillian, 1972; Beckwith et al., 1991) and a crucial building block of lexico-semantic networks and knowledge bases such as WordNet (Fellbaum, 1998), BabelNet (Navigli and Ponzetto, 2012) or ConceptNet (Speer et al., 2017). The ability to reason about concept-level entailment supports a plethora of tasks such as taxonomy induction (Snow et al., 2006; Navigli et al., 2011; Faralli et al., 2017), natural language inference (Dagan et al., 2013; Bowman et al., 2015; Williams et al., 2018), and text generation (Biran and McKeown, 2013) or metaphor detection (Mohler et al., 2013).

Binary and Graded Lexical Entailment. For this task, we follow the definition of lexical entailment as thoroughly discussed in Vulić et al. (2017, Section 2), namely as a taxonomical asymmetric hyponymy–hypernymy or *is-a* relation. Although commonly treated as a binary relation (*"Is X a type of Y?"*), cognitive theories of concept (proto)typicality and category vagueness (Rosch, 1975; Kamp and Partee, 1995) suggest that LE is rather a graded relation: humans can perceive the degree to which the LE relation holds between concepts (*"To which degree is X a type of Y?"*).[1] The graded nature of the LE relation has been empirically validated in human annotations crowdsourced for the **HyperLex** dataset (Vulić et al., 2017). Its creation catalyzed research on models for predicting graded LE (Nguyen et al., 2017; Nickel and Kiela, 2017; Vulić and Mrkšić, 2018; Tifrea et al., 2019; Le et al., 2019).

Multilingual and Cross-Lingual Lexical Entailment. Despite its potential for a variety of cross-lingual and multilingual applications such as multilingual taxonomy construction, machine translation, and multilingual natural language inference (Mihalcea et al., 2010; Negri et al., 2013; Ehrmann et al., 2014; Fu et al., 2014; Bordea et al., 2016; Conneau et al., 2018, *inter alia*), LE detection, especially its graded variant, has so far been predominantly studied in monolingual settings (Geffet and Dagan, 2005;

[1] For instance, *chess* is perceived as a less typical *sport* than *basketball*, but it is definitely a more typical *sport* than *sitting*.

Proceedings of the 14th International Workshop on Semantic Evaluation, pages 24–35
Barcelona, Spain (Online), December 12, 2020.

Weeds et al., 2014; Santus et al., 2014; Kiela et al., 2015; Shwartz et al., 2016; Shwartz et al., 2017; Glavaš and Ponzetto, 2017; Roller et al., 2018, *inter alia*) with most models and evaluations, unsurprisingly, focusing on English. Existing work on multilingual and cross-lingual LE (Vyas and Carpuat, 2016; Upadhyay et al., 2018; Glavaš and Vulić, 2019; Kamath et al., 2019) has been rather limited and focused dominantly on major and mutually similar languages and binary LE detection.

Shared Task. Aiming to catalyze the development of models for predicting LE, we organized the shared task described in this paper. Our shared task had a broad scope aiming to cover reasoning over lexical entailment from multiple perspectives. Namely, the subtasks covered both monolingual and cross-lingual setups as well as both binary LE detection and graded LE prediction (i.e., prediction of a degree to which LE holds between concepts). Our shared task encompassed a set of 6 etymologically and typologically diverse languages (and 15 corresponding language pairs): English (EN), German (DE), Italian (IT), Croatian (HR), Turkish (TR), and Albanian (SQ). We offered two different evaluation tracks. In the *distributional* (*Dist*) track we allowed only for fully distributional systems, capturing LE only on the basis of unannotated corpora. In contrast, the *Any* track invited systems that exploit any kind of additional external resources, including lexico-semantic networks.

Overall, we did not observe any empirically confirmed strong systems in the *Dist* track, further corroborating the findings from prior work that building LE-oriented vectors distributionally is more difficult than for some other relations such as broader semantic relatedness (Henderson and Popa, 2016). However, several runs submitted to the *Any* track pushed the state of the art both in binary LE detection and graded LE prediction, for most of the languages and language pairs in our evaluation.

2 Data

We started from the LE datasets we previously created and published (Vulić et al., 2017; Vulić et al., 2019b), covering four languages (EN, DE, IT, HR) and extended those datasets to two new languages (TR, SQ). For completeness, we describe the details of the annotation process and the creation of final multilingual and cross-lingual datasets for the shared task.

Starting Point: Graded LE in English. HyperLex (Vulić et al., 2017) comprises 2,616 English (EN) word pairs (2,163 noun pairs and 453 verb pairs) annotated for the graded LE relation. Unlike in symmetric similarity datasets (Hill et al., 2015; Gerz et al., 2016; Camacho-Collados et al., 2017), word order in each pair (X, Y) is important: this means that pairs (X, Y) and (Y, X) can obtain drastically different graded LE ratings. The word pairs were first sampled from WordNet to represent a spectrum of different word relations (e.g., hyponymy-hypernymy, meronymy, co-hyponymy, synonymy, antonymy, no relation). The ratings in the $[0, 6]$ interval were then collected through crowdsourcing by posing the graded LE *"To what degree is X a type of Y?"* question to human subjects, with each pair rated by at least 10 raters: the score of 6 indicates a perfect LE relation between the concepts X and Y (in that order), and 0 indicates absence of the LE relation. The final score was averaged across individual ratings. The final EN HyperLex dataset reveals that gradience effects are indeed present in human annotations: it contains word pairs with ratings distributed across the entire $[0, 6]$ rating interval.

Word Pair Translation. We first created monolingual HyperLex datasets in three target languages: German (DE), Italian (IT), and Croatian (HR), as described in (Vulić et al., 2019b). For this shared task, we repeated the procedure for two more languages: Turkish (TR), and our surprise test language – Albanian (SQ). We first translated word pairs from the EN HyperLex dataset and re-scored the translated pairs in the target language. The translation approach has been selected because (1) the original EN HyperLex pairs were already carefully selected through a controlled sampling procedure (ensuring a wide coverage of diverse relations). Moreover, (2) we wanted the datasets in different languages to be as comparable as possible in terms of concept coverage. The translation approach has been validated in previous work for creating multilingual semantic similarity datasets (Leviant and Reichart, 2015; Camacho-Collados et al., 2017). Most importantly, it allows for the automatic construction of cross-lingual graded LE datasets.

We have followed the standard word pair translation procedure (Leviant and Reichart, 2015; Camacho-Collados et al., 2017). Each EN HyperLex pair was first translated independently by two native speakers

Monolingual Examples				Cross-Lingual Examples			
EN	portrait	picture	5.90	EN-DE	dinosaur	Kreatur	4.75
EN	ascend	leave	1.08	EN-IT	eye	viso	0.6
EN	prestige	status	3.64	EN-HR	belief	religija	4.92
DE	Jazz	Musik	5.75	EN-SQ	harm	dëmtim	4.49
DE	Idol	Person	4.0	EN-TR	address	konuşmak	3.62
DE	erodieren	verschlechtern	0.25	DE-IT	Medikation	trattamento	5.38
IT	origano	cibo	3.25	DE-HR	Form	prizma	0.0
IT	sorella	comunità di donne	1.00	DE-SQ	Effekt	reaksion	1.43
IT	sfrattare	trasferire	2.75	DE-TR	verschwinden	solmak	0.71
HR	tenis	rekreacija	5.75	IT-HR	aritmetica	matematika	5.5
HR	iseliti	preseliti	3.00	IT-SQ	galleggiare	rrëshqas	0.6
SQ	biologji	shkencë	6.00	IT-TR	commedia	trajedi	0.0
SQ	rregulloj	bashkangjis	1.00	HR-SQ	živcirati	ngacmoj	2.45
TR	püskürmek	patlamak	1.67	HR-TR	terapija	tedavi	3.08
TR	tutmak	yakalamak	0.67	TR-SQ	alet	bisturi	5.55

Table 1: Example pairs with ratings from monolingual and cross-lingual graded LE datasets. Note: for cross-lingual datasets words from each language can be placed as either first or second in the pair.

	EN	DE	IT	HR	SQ	TR
EN	2116 / 1403	2521 / 2149	2841 / 2322	3016 / 2479	2638 / 2020	2616 / 2189
DE	–	2070 / 1681	2921 / 2614	3028 / 2745	2814 / 2305	2983 / 2662
IT	–	–	1977 / 1572	3170 / 2801	2757 / 2265	3044 / 2679
HR	–	–	–	2041 / 1645	2787 / 2344	3142 / 2765
SQ	–	–	–	–	1941 / 1235	2784 / 2281
TR	–	–	–	–	–	1972 / 1522

Table 2: Sizes of all monolingual (main diagonal) and cross-lingual LE test sets (format: graded/binary).

of the target language. We observed the translation agreement in the range of 80%-90% across the five target languages. Translation disagreements were resolved by a third annotator who selected the better of the two differing translations. We allowed for multi-word translations only if there was no appropriate single word translation, e.g., *typewriter* (EN) $\rightarrow$ *pisaći stroj* (HR).

Concept Scoring and Cross-Lingual Datasets. The resulting 2,616 concept pairs in all five target languages were annotated using a procedure analogous to that for EN HyperLex: the rating interval was $[0, 6]$, and each word pair was rated by 4 (for DE, IT, HR) or 5 (for TR, SQ) native speakers. We then constructed the cross-lingual datasets automatically, leveraging word pair translations and scores in five target languages. For this, we followed the established methodology, used for creating cross-lingual semantic similarity datasets (Camacho-Collados et al., 2015; Camacho-Collados et al., 2017). In short, we first intersect aligned concept pairs (obtained through translation) in two languages: e.g., *father-ancestor* in English and *otac-predak* in Croatian are used to create cross-lingual pairs *father-predak* and *otac-ancestor*. We then computed the graded LE scores of cross-lingual pairs as averages of corresponding monolingual scores. Finally, we retained only cross-lingual pairs for which the corresponding monolingual scores differ by ≤ 1.0: this heuristic (Camacho-Collados et al., 2017) mitigates the undesirable inter-language semantic shift. Table 1 shows examples of word pairs with score annotations from monolingual and cross-lingual datasets.

Final Shared Task Datasets. The obtained monolingual datasets slightly vary in size due to the elimination of unavoidable same-word pairs (i.e., the pairs in which both English words got translated into the same target language word). Cross-lingual datasets additionally vary in size because of the elimination of cross-lingual pairs for which the scores of the corresponding monolingual pairs mutually differed by more than 1.0. For each monolingual and cross-lingual dataset we separated 500 word pairs for the development portions and retained all remaining word pairs for the test portions used in the final evaluation.[2] For the

[2]We designated Albanian (SQ) to be a surprise language in our evaluation. Thus, we did not release the development portions of the monolingual SQ dataset nor the development portions of any of the five cross-lingual datasets involving Albanian.

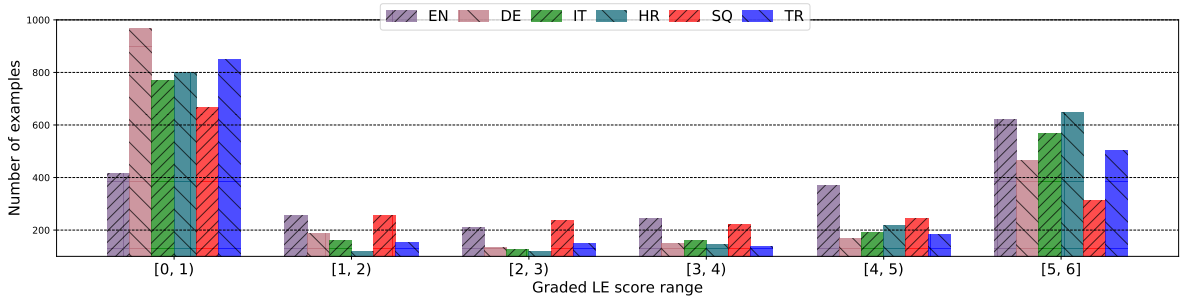

Figure 1: Score distributions (with 1-point wide buckets) for monolingual graded LE test sets.

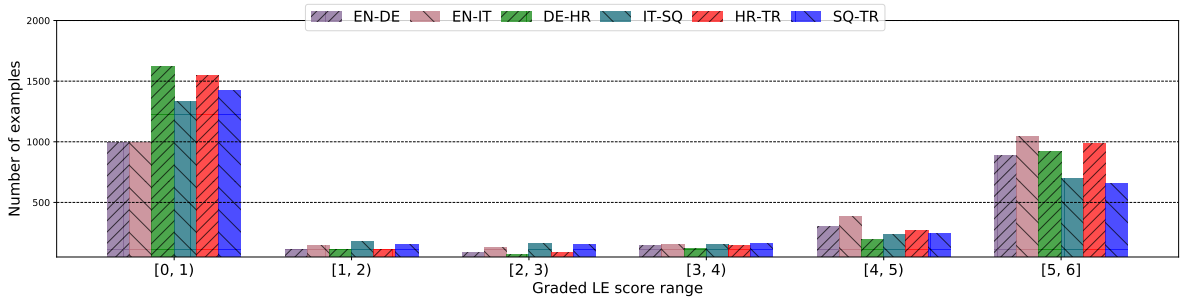

Figure 2: Score distributions for a sample of 6 cross-lingual graded LE test sets.

binary LE detection, we created a corresponding binary LE set from each graded LE dataset by (a) taking as positive examples all word pairs with the graded LE score ≥ 4.5, (b) taking as negative examples all word pairs with the graded LE score ≤ 1.5, and (c) eliminating word pairs with scores between 1.5 and 4.5. The sizes of all our final monolingual and cross-lingual test sets (both graded and binary) are shown in Table 2. The distribution of graded LE scores in monolingual test sets and (a sample of) cross-lingual test sets is given by Figure 1 and Figure 2, respectively. The majority of pairs are in the outer intervals (i.e., $[0, 1)$ and $[5, 6]$), with this being more pronounced for cross-lingual datasets. Nonetheless, the inner interval (i.e., $[1, 5)$) covers a significant portion ($\approx 30\%$) of (evenly distributed) word pairs, confirming the gradience of the LE relation.

3 Tasks, Subtasks, and Tracks

Tracks. The participants were asked to designate one of the two evaluation tracks for each of their submitted runs: (1) in the *Any* track we allowed for any kind of model/system to produce the LE predictions – the participants of this track were allowed to use external resources, including lexico-semantic networks like WordNet (Fellbaum, 1998) or BabelNet (Navigli and Ponzetto, 2012); (2) in contrast, in the *Dist* track, we allowed only for distributional models relying exclusively on unannotated corpora (of any size). We allowed each participant to submit at most 3 runs in each of the tracks.

Tasks and Subtasks. We defined four top-level tasks: (1) monolingual graded LE prediction, (2) monolingual binary LE detection, (3) cross-lingual graded LE prediction, and (4) cross-lingual binary LE detection. Each language in (1) and (2) (e.g., graded LE prediction for SQ) and each language-pair in (3) and (4) (e.g., binary LE detection for HR-TR) instantiates one concrete subtask. We allowed participants to submit their predictions for an arbitrary set of subtasks. Moreover, the participants were allowed to tackle only graded LE prediction or only binary LE detection.

Evaluation Metrics. For each graded LE prediction subtask, we measured the alignment of predictions and gold LE scores using the Spearman's Rank Correlation Coefficient (Spearman ρ), which is in line with previous work on similar concept pair scoring datasets (Hill et al., 2015; Levy et al., 2015; Vulić et

al., 2017, *inter alia*). For the binary LE detection subtasks we resorted to the standard F_1 measure.

4 Participating Systems

We now describe in more detail the approaches adopted by the three teams who submitted their system description papers.[3]

Team BMEAUT (Kovács et al., 2020). The BMEAUT method for LE detection and prediction is a rule-based approach that exploits Wiktionary definitions (Meyer and Gurevych, 2012) and relies on dependency parsing and semantic graphs. In the first step, the authors apply the `dict_to_4lang` tool (Recski et al., 2016) on Wiktionary definitions of concepts (which can be both unigrams and multi-word expressions, i.e., phrases) in order to induce the directed graphs conforming to the `4lang` formalism (Kornai et al., 2015). `4lang` graphs are directed graphs with concepts as nodes and three types of edges: edges of type 0 denote attribution (`cat` $\xrightarrow{0}$ `four-legged`), lexical entailment (`cat` $\xrightarrow{0}$ `mammal`), or unary predication (`cat` $\xrightarrow{0}$ `meow`); edges of type 1 and 2 denote relations between the predicate and its subject and object, respectively (e.g., `cat` $\xleftarrow{1}$ `catch` $\xrightarrow{2}$ `mouse`).[4]

Kovács et al. (2020) first extract definitions from Wiktionary using language-specific templates. Each definition is then transformed into a `4lang` graph with the help of a language-specific Universal Dependencies (Nivre et al., 2016) parser. Let (x, y) be the candidate word pair from one of our test sets. Kovács et al. (2020) then start from the `4lang` graph of x and include all concepts to which x has an outgoing edge of type 0. They then iteratively expand this initial `4lang` graph of x by adding in each iteration all concepts for which they have an incoming edge of type 0 from any of the nodes already in the graph. If, via this extension, they get to include y into the graph, they conclude (in the binary setup) that x indeed entails y, i.e., that (x, y) is a positive LE instance. Somewhat expectedly, this rule-based approach has a very high precision, but relatively low recall. This is why, in the next step, the authors augment their `4lang` graph containing extractions from Wiktionary definitions with the hypernymy pairs from WordNet (for EN and IT they use language-specific WordNets, for DE they translate German terms to English and use English WordNet).

The reliance on time-consuming rule-based design of language-specific Wiktionary extractors and language-specific LE detection rules on `4lang` graphs prevented the authors from submitting predictions for lower-resource languages (HR, SQ, TR). Also, the authors did not submit any graded LE predictions nor cross-lingual predictions. Building reliable scores on a continuous scale has been proven difficult for semantic similarity (Wu and Palmer, 1994; Lin and others, 1998) and distributional similarity measures have been shown to perform better. We presume something similar to be the case with LE and the proposed `4lang` graphs – it is inherently difficult to create a reliable LE score based on paths and distances in a symbolic representation that is a (directed) graph.

Team UAlberta (Hauer et al., 2020). The approach of UAlberta for cross-lingual binary LE detection combines sentence-level translations (i.e., parallel corpora), distributional word vectors (i.e., word embeddings) and multilingual lexical resources. Their base method, dubbed BITEXT, mines candidates for cross-lingual LE from parallel corpora – they simply run the FastAlign (Dyer et al., 2013) word alignment algorithm and assume that the LE relation holds between all aligned pairs of words. As clarified by the authors, this will, in most cases, extract cross-lingual synonyms, which, strictly speaking, do satisfy the LE relation; also, in some cases, the alignments will be established between close (e.g., first order) hyponymy-hypernymy pairs – in this case, however, the bitext alignment of words alone does not suggest the direction of the LE relation. The authors simply declare any pair of words from our cross-lingual datasets to stand in the LE relation if they find this pair in the list of generated word alignments. The second run, dubbed VECTORS, extends BITEXT by exploiting similarities between distributional embeddings of words. To paraphrase the authors, the intuition behind VECTORS is *that mutually semantically similar*

[3] We received partial run submissions from 3 other teams. However, these teams never sent us a short methodological description nor did they submit a system description paper.

[4] In that sense, edges of type 1 and 2 very much correspond to standard dependency relations of *nominal subject* and *direct object*, respectively.

concepts tend to entail the same set of other concepts. Let (x, y) be a cross-lingual word alignment generated with BITEXT. Let $X = \{x_1, \ldots, x_n\}$ be the set of terms which are semantically most similar to x in one language and $Y = \{y_1, \ldots, y_m\}$ be the set of terms which are semantically most similar to y in the other language. Note that this does not require a bilingual word embedding space, merely two monolingual word embedding spaces: sets X and Y are obtained by thresholding monolingual word embedding similarities (the threshold value is tuned on the development portions of our LE sets). Finally, all possible pairs $(x_i, y_j) \in X \times Y$ are considered to stand in the LE relation.

Finally, in the third run, the authors couple the bitext-based FastText aligner with the BABELALIGN algorithm, which aligns concepts across languages based on BabelNet (Navigli and Ponzetto, 2012), a massively multilingual lexico-semantic network.

Team SHIKEBLCU (Wang et al., 2020). The approach of Wang et al. (2020) extends the well-established line of work based on specializing (i.e., fine-tuning) distributional word vectors for lexical relations, be it symmetric semantic similarity (Faruqui et al., 2015; Mrkšić et al., 2017; Vulić et al., 2018; Glavaš and Vulić, 2018; Ponti et al., 2018) or the asymmetric LE relation (Vulić and Mrkšić, 2018; Glavaš and Vulić, 2019; Kamath et al., 2019; Vulić et al., 2019b), using constraints from external lexico-semantic resources like WordNet for supervision. At the core of the approach is the Lexical-Entailment Attract Repel (LEAR) (Vulić and Mrkšić, 2018), a retrofitting model that specializes distributional vectors of words from external constraints (synonyms, antonyms, and LE pairs). Wang et al. (2020) first LE-specialize with LEAR monolingual word embedding spaces of each language independently using language-specific constraints collected from ConceptNet (Speer et al., 2017). The number of constraints they obtain for other languages is, however, significantly smaller than for EN, with especially few constraints obtained for lower-resource languages – HR, TR, and SQ. Because of this, they add additional constraints in target languages by translating EN constraints to target languages via Google Translate. A similar approach of automatic constraint translation has already been proven very effective in the context of symmetric similarity-based specialization of embedding spaces for low-resource languages (Ponti et al., 2019). This way, Wang et al. (2020) obtain an LE-specialized embedding space for each language.

Following that, in the second step, they learn a linear projection mapping between the LE-specialized monolingual spaces with the VecMap tool for inducing bilingual word embedding spaces (Artetxe et al., 2018). Recent comparative evaluations (Glavaš et al., 2019; Vulić et al., 2019a) rendered VecMap as one of the most robust algorithms for inducing cross-lingual embedding spaces. The word translations obtained with Google Translate when translating EN constraints are also forwarded to VecMap as supervision for inducing bilingual embedding spaces.

5 Official Evaluation

We now report the official results of our evaluation. We first describe the baselines (Section 5.1) and then show the performances for all submitted runs (Section 5.2).[5]

5.1 Baselines

For the *Dist* track we use simple cosine similarity between distributional word vectors as a baseline. To this end, we use the 300-dimensional FastText embeddings (Bojanowski et al., 2017) trained on Wikipedias of respective languages.[6] For the cross-lingual (sub)tasks we induce the bilingual embedding spaces via the simple Procrustes alignment (Smith et al., 2017), using 5K word translation dictionaries, as described in Glavaš et al. (2019). Since LE is an asymmetric relation and cosine similarity is a symmetric measure, we did not expect this baseline to be particularly competitive and expected most participants to outperform it.

For the *Any* track, we used GLEN, our recent neural *explicit specialization* model for LE (Glavaš and Vulić, 2019) as a competitive baseline. GLEN is a simple feed-forward network that learns to specialize distributional word vectors for LE based on three types of lexico-semantic constraints (originating primarily from WordNet): synonyms, antonyms, and LE constraints. Unlike the standard retrofitting models (Faruqui et al., 2015; Mrkšić et al., 2017; Vulić and Mrkšić, 2018), in explicit retrofitting the constraints are not

[5] For completeness, we also report the scores for runs for which we have received no system description papers.

[6] `https://fasttext.cc/docs/en/pretrained-vectors.html`

Submission	EN	DE	IT	HR	TR	SQ
Distributional models (DIST track)						
Baseline: Cosine	77.80	52.01	63.27	64.52	58.0	54.11
AYAH (run 1)	74.63	43.42	54.01	–	46.54	–
SPON (run 1)	75.27	–	–	–	–	–
Any resource models (ANY track)						
Baseline: GLEN	79.87	59.88	66.27	64.27	64.35	56.86
BMEAUT (run 1)	**91.77**	67.00	**81.41**	–	–	–
FERRYMAN (run 1)	72.13	53.12	62.71	55.62	38.45	38.67
FERRYMAN (run 2)	72.13	53.2	62.71	55.62	38.45	38.67
SHIKEBLCU (run 1)	87.90	**71.43**	75.94	**75.37**	**69.85**	**72.12**

Table 3: Monolingual LE results for **binary** LE detection (F_1 scores).

Submission	EN	DE	IT	HR	TR	SQ
Distributional models (DIST track)						
Baseline: Cosine	18.59	22.02	12.42	15.5	16.54	26.13
Any resource models (ANY track)						
Baseline: GLEN	51.24	43.31	43.2	38.29	43.06	32.25
FERRYMAN (run 2)	34.62	27.59	35.39	24.56	15.39	9.66
SHIKEBLCU (run 1)	**69.63**	**63.17**	**63.6**	**58.85**	**52.52**	**56.45**

Table 4: Monolingual LE results for **graded** LE prediction (Spearman ρ).

used to directly tune the vectors of the words from those constraints, but are rather exploited as training examples to learn a general specialization function (in case of GLEN, a feed-forward network). This way, explicit retrofitting models (Glavaš and Ponzetto, 2017; Glavaš and Vulić, 2018; Glavaš and Vulić, 2019) can specialize any distributional vector from the original monolingual space and not just the vectors of words from the constraints, as is the case with classic retrofitting models. We use the pre-trained GLEN instance from our original work, which was trained using only EN constraints. In order to make GLEN applicable in monolingual tasks in other languages as well as in the cross-lingual tasks, we first project monolingual embeddings of other languages to the EN monolingual embedding space. For more details, we refer the reader to the original paper (Glavaš and Vulić, 2019).

Both baselines (cosine similarity in the *Dist* track and GLEN in the *Any* track) produce real-valued scores which can be directly used for our graded LE evaluations. For the binary LE detection subtasks, we first binarize the scores on some threshold value: for both baselines we tune the threshold value on respective development dataset portions.

5.2 Results

Monolingual Results. In Table 3 we report the results for all monolingual binary LE detection subtasks. Table 4 displays the results for the respective graded LE subtasks. Unfortunately, we have not seen any encouraging results in the *Dist* track: the two runs submitted for the binary LE detection fail to outperform the simple cosine similarity baseline;[7] and we have not received any submissions for the graded LE prediction for the *Dist* track. Despite the limited number of overall submissions, this also points to the complexity of graded LE reasoning based solely on raw text data and distributional signal.

The results in the *Any* track are, however, much more exciting. Both Wang et al. (2020) and Kovács et al. (2020) manage to outperform our competitive baseline GLEN in binary LE detection, in some cases by a fairly wide margin (e.g., BMEAUT for EN and IT, SHIKEBLCU for HR and SQ). The results achieved by SHIKEBLCU in the graded monolingual LE tasks (Table 4) are even more encouraging and truly push the state-of-the-art in graded multilingual LE prediction – the improvements over GLEN are ≥ 20 Spearman correlation points for low-resource languages in our evaluation (TR, HR, SQ). Both BMEAUT and SHIKELBCU use language-specific constraints (BMEAUT by extracting 4LANG graphs from language-specific Wiktionaries and SHIKELBCU by translating the EN constraints from WordNet),

[7] We have not received any system descriptions for these runs.

Submission	de-en	de-hr	de-it	de-sq	de-tr	en-hr	en-it	en-sq	en-tr	hr-it	hr-sq	hr-tr	it-sq	it-tr	sq-tr
Distributional models (DIST track)															
Baseline: Cosine	71.7	57.5	55.6	51.1	54.3	74.0	74.3	66.2	72.3	63.2	57.3	61.5	58.4	60.3	55.3
AYAH (run 1)	62.4	–	46.5	–	44.4	42.2	67.8	–	63.7	–	–	–	–	50.5	–
Any resources models (ANY track)															
Baseline: GLEN	74.3	62.6	63.7	58.8	**63.2**	65.9	77.2	65.7	74.3	61.6	57.6	63.7	59.8	**67.6**	61.2
FERRYMAN (run 1)	66.1	55.7	59.5	46.2	47.1	67.6	71.5	56.9	59.2	60.5	48.8	49.2	51.9	52.6	35.9
FERRYMAN (run 2)	66.1	55.7	59.5	46.2	47.1	67.6	71.5	56.9	59.2	60.5	48.8	49.2	51.9	52.6	35.9
SHIKEBLCU (run 1)	**80.6**	**64.0**	**63.8**	**61.4**	62.7	**78.8**	**81.4**	**74.8**	**77.8**	**69.0**	**63.6**	**65.0**	**67.0**	67.5	**62.2**
UALBERTA (run 1)	63.1	47.6	49.6	43.2	46.4	64.2	67.9	52.0	61.2	55.4	46.4	44.4	50.7	52.2	43.8
UALBERTA (run 2)	65.0	54.7	51.7	–	–	–	74.3	–	–	–	–	–	–	–	–
UALBERTA (run 3)	70.7	55.5	56.6	–	–	–	75.3	–	–	–	–	–	–	–	–

Table 5: Cross-lingual LE results for **binary** LE prediction (F_1 scores).

Submission	de-en	de-hr	de-it	de-sq	de-tr	en-hr	en-it	en-sq	en-tr	hr-it	hr-sq	hr-tr	it-sq	it-tr	sq-tr
Distributional models (DIST track)															
Baseline: Cosine	30.5	24.3	22.5	33.5	25.9	26.8	27.1	34.6	34.2	20.1	33.6	23.1	35.2	22.3	36.4
Any resources models (ANY track)															
Baseline: GLEN	50.4	40.8	**48.8**	39.6	**46.6**	36.8	**54.1**	39.3	50.4	35.3	36.6	**40.2**	35.4	**47.4**	37.2
FERRYMAN (run 2)	39.8	30.6	36.3	22.1	21.9	37.6	42.7	26.9	27.2	32.6	18.9	21.9	24.3	23.8	9.8
SHIKEBLCU (run 1)	**56.7**	**42.1**	45.7	**52.9**	45.8	**49.4**	53.2	**55.9**	**51.0**	**43.2**	**49.0**	38.6	**54.4**	45.9	**51.0**

Table 6: Cross-lingual LE results for **graded** LE prediction (Spearman ρ scores).

whereas GLEN relies only on EN constraints and then transfers the LE specialization to other languages via cross-lingual word embeddings. The especially good performance of SHIKEBLCU – who translate large constraint sets to target languages – in graded LE is aligned with a similar finding established for the symmetric relation of semantic similarity: (a) translation of constraints and specialization directly in the target language (Ponti et al., 2019) substantially outperforms (b) the specialization in the source language (EN) followed by a transfer of the specialization model to the target languages via cross-lingual embedding spaces (Vulić et al., 2018; Glavaš and Vulić, 2018).

Cross-Lingual Results. The results of our cross-lingual evaluation are shown in Table 5 (binary LE detection) and Table 6 (graded LE prediction). The results mostly follow the same trends of the monolingual results: we received no successful runs in the *Dist* track, but we see very encouraging results in the *Any* track. Similar to the monolingual settings, SHIKEBLCU outperforms the competitive GLEN baseline, although now with somewhat narrower margins and not for all language pairs, especially in the graded LE setup (Table 6). An encouraging finding is that largest gains with SHIKEBLCU over GLEN are for language pairs involving Albanian (SQ), our surprise evaluation language and arguably the most resource-lean language in our evaluation (e.g., the margins in graded LE prediction in favor of SHIKEBLCU are 13%, 19%, and 14% for HR-SQ, IT-SQ, and SQ-TR, respectively). Although none of the UAlberta runs (Hauer et al., 2020) outperform GLEN, it is encouraging to see that competitive performance can be achieved by simple methods with relatively low-resource demands (e.g., their run-1, which is their VECTORS approach, requires only parallel corpora and monolingual word embedding spaces).

6 Conclusion

As a fundamental asymmetric lexico-semantic relation, lexical entailment (LE) supports construction of concept hierarchies and downstream applications that require reasoning and inference. In multilingual and cross-lingual applications we need models that can detect LE across languages. This is why we carried out a SemEval shared task (Task 2) on predicting LE, spanning (1) monolingual vs. cross-lingual LE detection, (2) binary LE detection vs. graded LE prediction, and (3) a set of 6 diverse languages (and 15 language pairs). In the track in which we allowed for any external resources to be used (*Any* track), we received submissions that substantially push the state of the art across all languages and language pairs, for both binary LE detection and graded LE prediction. We hope that these methodological advances, instigated

by the SemEval task and the constructed datasets, will inform and inspire further work in fields such as multilingual taxonomy induction and language inference.

Acknowledgements

We thank all participants for submitting their runs, despite (to put it mildly) less than ideal and extremely uncertain circumstances caused by the COVID-19 pandemic. We thank our annotators, native speakers of German, Italian, Croatian, Turkish, and Albanian for their high-quality and dedicated work on creating datasets for this shared task. Goran Glavaš is supported by the Baden-Württemberg Stiftung (Eliterprogramm, grant AGREE). Simone Paolo Ponzetto is supported by the German Science Foundation (JOIN-T 2 grant). Ivan Vulić and Anna Korhonen are supported by the ERC Consolidator Grant LEXICAL (no. 648909).

References

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of ACL*, pages 789–798.

Richard Beckwith, Christiane Fellbaum, Derek Gross, and George A. Miller. 1991. WordNet: A lexical database organized on psycholinguistic principles. *Lexical acquisition: Exploiting on-line resources to build a lexicon*, pages 211–231.

Or Biran and Kathleen McKeown. 2013. Classifying taxonomic relations between pairs of Wikipedia articles. In *Proceedings of IJCNLP*, pages 788–794.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the ACL*, 5:135–146.

Georgeta Bordea, Els Lefever, and Paul Buitelaar. 2016. Semeval-2016 task 13: Taxonomy extraction evaluation (TExEval-2). In *Proceedings of SEMEVAL*, pages 1081–1091.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of EMNLP*, pages 632–642.

José Camacho-Collados, Mohammad Taher Pilehvar, and Roberto Navigli. 2015. A framework for the construction of monolingual and cross-lingual word similarity datasets. In *Proceedings of ACL*, pages 1–7.

Jose Camacho-Collados, Mohammad Taher Pilehvar, Nigel Collier, and Roberto Navigli. 2017. Semeval-2017 task 2: Multilingual and cross-lingual semantic word similarity. In *Proceedings of SEMEVAL*, pages 15–26.

Allan M. Collins and Ross M. Quillian. 1972. Experiments on semantic memory and language comprehension. *Cognition in Learning and Memory*.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of EMNLP*, pages 2475–2485.

Ido Dagan, Dan Roth, Mark Sammons, and Fabio Massimo Zanzotto. 2013. Recognizing textual entailment: Models and applications. *Synthesis Lectures on Human Language Technologies*, 6(4):1–220.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of IBM Model 2. In *Proceedings of NAACL-HLT*, pages 644–648.

Maud Ehrmann, Francesco Cecconi, Daniele Vannella, John Philip McCrae, Philipp Cimiano, and Roberto Navigli. 2014. Representing multilingual data as linked data: the case of BabelNet 2.0. In *Proceedings of LREC*, pages 401–408.

Stefano Faralli, Alexander Panchenko, Chris Biemann, and Simone Paolo Ponzetto. 2017. The ContrastMedium algorithm: Taxonomy induction from noisy knowledge graphs with just a few links. In *Proceedings of EACL*, pages 590–600.

Manaal Faruqui, Jesse Dodge, Sujay Kumar Jauhar, Chris Dyer, Eduard Hovy, and Noah A. Smith. 2015. Retrofitting word vectors to semantic lexicons. In *Proceedings of NAACL-HLT*, pages 1606–1615.

Christiane Fellbaum. 1998. *WordNet*. MIT Press.

Ruiji Fu, Jiang Guo, Bing Qin, Wanxiang Che, Haifeng Wang, and Ting Liu. 2014. Learning semantic hierarchies via word embeddings. In *Proceedings of ACL*, pages 1199–1209.

Maayan Geffet and Ido Dagan. 2005. The distributional inclusion hypotheses and lexical entailment. In *Proceedings of ACL*, pages 107–114.

Daniela Gerz, Ivan Vulić, Felix Hill, Roi Reichart, and Anna Korhonen. 2016. SimVerb-3500: A large-scale evaluation set of verb similarity. In *Proceedings of EMNLP*, pages 2173–2182.

Goran Glavaš and Ivan Vulić. 2018. Explicit retrofitting of distributional word vectors. In *Proceedings of ACL*, pages 34–45.

Goran Glavaš and Ivan Vulić. 2019. Generalized tuning of distributional word vectors for monolingual and cross-lingual lexical entailment. In *Proceedings of ACL*, pages 4824–4830.

Goran Glavaš, Robert Litschko, Sebastian Ruder, and Ivan Vulić. 2019. How to (properly) evaluate cross-lingual word embeddings: On strong baselines, comparative analyses, and some misconceptions. In *Proceedings of ACL*, pages 710–721.

Goran Glavaš and Simone Paolo Ponzetto. 2017. Dual tensor model for detecting asymmetric lexico-semantic relations. In *Proceedings of EMNLP*, pages 1758–1768.

Bradley Hauer, Amir Ahmad Habibi, Yixing Luan, Arnob Mallik, and Grzegorz Kondrak. 2020. UAlberta at SemEval-2020 Task 2: Using translations to predict cross-lingual entailment. In *Proceedings of SEMEVAL*.

James Henderson and Diana Nicoleta Popa. 2016. A vector space for distributional semantics for entailment. In *Proceedings of ACL*, pages 2052–2062.

Felix Hill, Roi Reichart, and Anna Korhonen. 2015. SimLex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41(4):665–695.

Aishwarya Kamath, Jonas Pfeiffer, Edoardo Maria Ponti, Goran Glavaš, and Ivan Vulić. 2019. Specializing distributional vectors of all words for lexical entailment. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, pages 72–83.

Hans Kamp and Barbara Partee. 1995. Prototype theory and compositionality. *Cognition*, 57(2):129–191.

Douwe Kiela, Laura Rimell, Ivan Vulić, and Stephen Clark. 2015. Exploiting image generality for lexical entailment detection. In *Proceedings of ACL*, pages 119–124.

András Kornai, Judit Acs, Márton Makrai, Dávid Márk Nemeskey, Katalin Pajkossy, and Gábor András Recski. 2015. Competence in lexical semantics. In *Proceedings of *SEM*, pages 165–175.

Ádám Kovács, Kinga Gémes, András Kornai, and Gábor Recski. 2020. BMEAUT at SemEval-2020 Task 2: Lexical entailment with semantic graphs. In *Proceedings of SEMEVAL*.

Matt Le, Stephen Roller, Laetitia Papaxanthos, Douwe Kiela, and Maximilian Nickel. 2019. Inferring concept hierarchies from text corpora via hyperbolic embeddings. In *Proceedings of ACL*, pages 3231–3241.

Ira Leviant and Roi Reichart. 2015. Separated by an un-common language: Towards judgment language informed vector space modeling. *CoRR*, abs/1508.00106.

Omer Levy, Yoav Goldberg, and Ido Dagan. 2015. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the ACL*, 3:211–225.

Dekang Lin et al. 1998. An information-theoretic definition of similarity. In *Proceedings of ICML*, pages 296–304.

Christian M. Meyer and Iryna Gurevych. 2012. Wiktionary: A new rival for expert-built lexicons? exploring the possibilities of collaborative lexicography. In *Electronic Lexicography*, pages 259–291.

Rada Mihalcea, Ravi Som Sinha, and Diana McCarthy. 2010. Semeval-2010 task 2: Cross-lingual lexical substitution. In *Proceedings of SEMEVAL*, pages 9–14.

Michael Mohler, David Bracewell, Marc Tomlinson, and David Hinote. 2013. Semantic signatures for example-based linguistic metaphor detection. In *Proceedings of the First Workshop on Metaphor in NLP*, pages 27–35.

Nikola Mrkšić, Ivan Vulić, Diarmuid Ó Séaghdha, Ira Leviant, Roi Reichart, Milica Gašić, Anna Korhonen, and Steve Young. 2017. Semantic specialisation of distributional word vector spaces using monolingual and cross-lingual constraints. *Transactions of the ACL*, 5:309–324.

Roberto Navigli and Simone Paolo Ponzetto. 2012. Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

Roberto Navigli, Paola Velardi, and Stefano Faralli. 2011. A graph-based algorithm for inducing lexical taxonomies from scratch. In *Proceedings of IJCAI*, pages 1872–1877.

Matteo Negri, Alessandro Marchetti, Yashar Mehdad, Luisa Bentivogli, and Danilo Giampiccolo. 2013. Semeval-2013 task 8: Cross-lingual textual entailment for content synchronization. In *Proceedings of SEMEVAL*, pages 25–33.

Kim Anh Nguyen, Maximilian Köper, Sabine Schulte im Walde, and Ngoc Thang Vu. 2017. Hierarchical embeddings for hypernymy detection and directionality. In *Proceedings of EMNLP*, pages 233–243.

Maximilian Nickel and Douwe Kiela. 2017. Poincaré embeddings for learning hierarchical representations. In *Proceedings of NeurIPS*, pages 6341–6350.

Joakim Nivre, Marie-Catherine De Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, et al. 2016. Universal Dependencies v1: A multilingual treebank collection. In *Proceedings of LREC*, pages 1659–1666.

Edoardo Maria Ponti, Ivan Vulić, Goran Glavaš, Nikola Mrkšić, and Anna Korhonen. 2018. Adversarial propagation and zero-shot cross-lingual transfer of word vector specialization. In *Proceedings of EMNLP*, pages 282–293.

Edoardo Maria Ponti, Ivan Vulić, Goran Glavaš, Roi Reichart, and Anna Korhonen. 2019. Cross-lingual semantic specialization via lexical relation induction. In *Proceedings of EMNLP-IJCNLP*, pages 2206–2217.

Gábor Recski, Eszter Iklódi, Katalin Pajkossy, and Andras Kornai. 2016. Measuring semantic similarity of words using concept networks. In *Proceedings of the 1st Workshop on Representation Learning for NLP*, pages 193–200.

Stephen Roller, Douwe Kiela, and Maximilian Nickel. 2018. Hearst patterns revisited: Automatic hypernym detection from large text corpora. In *Proceedings of ACL*, pages 358–363.

Eleanor H. Rosch. 1975. Cognitive representations of semantic categories. *Journal of Experimental Psychology*, 104(3):192–233.

Enrico Santus, Alessandro Lenci, Qin Lu, and Sabine Schulte im Walde. 2014. Chasing hypernyms in vector spaces with entropy. In *Proceedings of EACL*, pages 38–42.

Vered Shwartz, Yoav Goldberg, and Ido Dagan. 2016. Improving hypernymy detection with an integrated path-based and distributional method. In *Proceedings of ACL*, pages 2389–2398.

Vered Shwartz, Enrico Santus, and Dominik Schlechtweg. 2017. Hypernyms under siege: Linguistically-motivated artillery for hypernymy detection. In *Proceedings of EACL*, pages 65–75.

Samuel L. Smith, David H.P. Turban, Steven Hamblin, and Nils Y. Hammerla. 2017. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. In *Proceedings of ICLR*.

Rion Snow, Daniel Jurafsky, and Andrew Y. Ng. 2006. Semantic taxonomy induction from heterogenous evidence. In *Proceedings of ACL*, pages 801–808.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. ConceptNet 5.5: An open multilingual graph of general knowledge. In *Proceedings of AAAI*, pages 4444–4451.

Alexandru Tifrea, Gary Bécigneul, and Octavian-Eugen Ganea. 2019. Poincaré GloVe: Hyperbolic word embeddings. In *Proceedings of ICLR*.

Shyam Upadhyay, Yogarshi Vyas, Marine Carpuat, and Dan Roth. 2018. Robust cross-lingual hypernymy detection using dependency context. In *Proceedings of NAACL-HLT*, pages 607–618.

Ivan Vulić and Nikola Mrkšić. 2018. Specialising word vectors for lexical entailment. In *Proceedings of NAACL-HLT*, pages 1134–1145.

Ivan Vulić, Daniela Gerz, Douwe Kiela, Felix Hill, and Anna Korhonen. 2017. Hyperlex: A large-scale evaluation of graded lexical entailment. *Computational Linguistics*, 43(4):781–835.

Ivan Vulić, Goran Glavaš, Nikola Mrkšić, and Anna Korhonen. 2018. Post-specialisation: Retrofitting vectors of words unseen in lexical resources. In *Proceedings of NAACL-HLT*, pages 516–527.

Ivan Vulić, Goran Glavaš, Roi Reichart, and Anna Korhonen. 2019a. Do we really need fully unsupervised cross-lingual embeddings? In *Proceedings EMNLP-IJCNLP*, pages 4398–4409.

Ivan Vulić, Simone Paolo Ponzetto, and Goran Glavaš. 2019b. Multilingual and cross-lingual graded lexical entailment. In *Proceedings of ACL*, pages 4963–4974.

Yogarshi Vyas and Marine Carpuat. 2016. Sparse bilingual word representations for cross-lingual lexical entailment. In *Proceedings of NAACL-HLT*, pages 1187–1197.

Shike Wang, Yuchen Fan, Xiangying Luo, and Dong Yu. 2020. SHIKEBLCU at SemEval-2020 Task 2: An external knowledge-enhanced matrix for multilingual and cross-lingual lexical entailment. In *Proceedings of SEMEVAL*.

Julie Weeds, Daoud Clarke, Jeremy Reffin, David Weir, and Bill Keller. 2014. Learning to distinguish hypernyms and co-hyponyms. In *Proceedings of COLING*, pages 2249–2259.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of NAACL-HLT*, pages 1112–1122.

Zhibiao Wu and Martha Palmer. 1994. Verbs semantics and lexical selection. In *Proceedings of ACL*, pages 133–138.