Multilingual and Cross-Lingual Graded Lexical Entailment

Ivan Vulić[1], **Simone Paolo Ponzetto**[2], **Goran Glavaš**[2]
[1] PolyAI Ltd., London, United Kingdom
[2] Data and Web Science Group, University of Mannheim, Germany
ivan@poly-ai.com
{simone, goran}@informatik.uni-mannheim.de

Abstract

Grounded in cognitive linguistics, graded lexical entailment (GR-LE) is concerned with fine-grained assertions regarding the directional hierarchical relationships between concepts on a continuous scale. In this paper, we present the first work on cross-lingual generalisation of GR-LE relation. Starting from HyperLex, the only available GR-LE dataset in English, we construct new monolingual GR-LE datasets for three other languages, and combine those to create a set of six cross-lingual GR-LE datasets termed CL-HYPERLEX. We next present a novel method dubbed CLEAR (**C**ross-**L**ingual **L**exical **E**ntailment **A**ttract-**R**epel) for effectively capturing graded (and binary) LE, both monolingually in different languages as well as across languages (i.e., on CL-HYPERLEX). Coupled with a bilingual dictionary, CLEAR leverages taxonomic LE knowledge in a resource-rich language (e.g., English) and propagates it to other languages. Supported by cross-lingual LE transfer, CLEAR sets competitive baseline performance on three new monolingual GR-LE datasets and six cross-lingual GR-LE datasets. In addition, we show that CLEAR outperforms current state-of-the-art on binary cross-lingual LE detection by a wide margin for diverse language pairs.

1 Introduction

Word-level lexical entailment (LE), also known as the TYPE-OF or hyponymy-hypernymy relation, is a fundamental *asymmetric* lexical relation (Collins and Quillian, 1972; Beckwith et al., 1991). It is a key principle behind the hierarchical structure found in semantic networks such as WordNet (Fellbaum, 1998) or ConceptNet (Speer et al., 2017).

As opposed to simpler discrete and binary LE detection (e.g., *oregano* is a TYPE-OF *food*), *graded* lexical entailment (GR-LE) measures the strength of the LE relation between two concepts on a continuous scale (Vulić et al., 2017; Rei et al., 2018). GR-LE is concerned with fine-grained directional assertions of hierarchical arrangements between concepts. The notion of graded LE is rooted in theories of concept (proto)typicality and category vagueness from cognitive science (Rosch, 1973, 1975; Kamp and Partee, 1995). Instead of answering the simpler (discrete) question *"Is X a type of Y?"*, as in standard LE detection tasks (Kotlerman et al., 2010; Turney and Mohammad, 2015), GR-LE aims at answering the following question: *"To what degree is X a type of Y?"* The concept of LE gradience is also empirically confirmed by human judgements elicited for HyperLex (Vulić et al., 2017), a GR-LE resource in English.[1]

Furthermore, while simpler binary LE detection has been predominantly studied in monolingual settings only (Geffet and Dagan, 2005; Weeds et al., 2014; Santus et al., 2014; Kiela et al., 2015; Shwartz et al., 2016, 2017; Glavaš and Ponzetto, 2017; Roller et al., 2018, *inter alia*), more general reasoning over cross-lingual and multilingual LE relationships can improve language understanding in multilingual contexts, e.g., in cases when translations are ambiguous or not equivalent to the source concept (Vyas and Carpuat, 2016; Upadhyay et al., 2018).[2] The ability to reason over cross-lingual LE is pivotal for a variety of cross-lingual tasks such as recognising cross-lingual textual entailment (Negri et al., 2012, 2013; Conneau et al., 2018b), constructing multilingual taxonomies (Ehrmann et al., 2014; Fu et al., 2014), cross-lingual event coreference (Song et al., 2018), machine translation in-

[1] For instance, the strength of LE association *hamburger* → *food* is on average judged by humans with 5.85/60. In comparison, *oregano* is seen as a less typical instance of the category/concept *food*, with the pair's average rating of 3.58/6.0. In contrast, the pair *food* → *pie* receives the average rating of only 0.92/6, which confirms the inherent asymmetry of the GR-LE relation.

[2] For instance, translating the Italian word *calcio* to *calcium* prevents identifying *sport* as a hypernym of *calcio*.

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 4963–4974
Florence, Italy, July 28 - August 2, 2019.

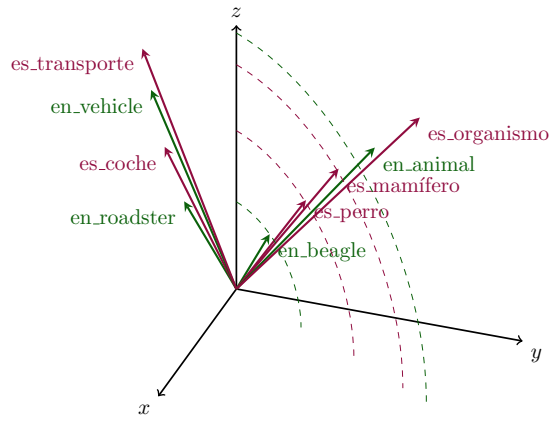


Figure 1: A toy example of Euclidean shared cross-lingual word vector space specialised for the asymmetric LE relation. The symmetric similarity of true LE pairs, irrespective of their actual language (the example shows English and Spanish words with the respective prefixes *en_* and *es_*) is reflected by their small cosine distances (e.g., the small angle between $\overrightarrow{en\_beagle}$ and $\overrightarrow{es\_perro}$ and $\overrightarrow{en\_animal}$), while simultaneously higher-level concepts are assigned larger norms to enforce the LE arrangement in the vector space. An asymmetric distance that takes into account the vector direction as well as the vector magnitude can be used to grade the LE relation strength between any two concepts in the shared cross-lingual vector space.

terpretability (Padó et al., 2009), and cross-lingual lexical substitution (Mihalcea et al., 2010).

In this work, we introduce the first set of benchmarks and methods that target cross-lingual and multilingual graded lexical entailment. We make several important contributions related to GR-LE in multilingual settings. First, we extend the research on GR-LE beyond English (Vulić et al., 2017; Rei et al., 2018) and provide new human-annotated GR-LE datasets in three other languages: German, Italian, and Croatian. Second, following an established methodology for constructing evaluation datasets for cross-lingual lexico-semantic relations (Camacho-Collados et al., 2015, 2017), we automatically derive a collection of six cross-lingual GR-LE datasets: CL-HYPERLEX. We analyse in detail the cross-lingual datasets (e.g., by comparing the scores to human-elicited ratings), demonstrating their robustness and reliability.

In order to provide a competitive baseline on new monolingual and cross-lingual datasets, we next introduce a cross-lingual specialisation/retrofitting method termed CLEAR (**C**ross-**L**ingual **L**exical **E**ntailment **A**ttract-**R**epel): starting from *any* two monolingual distributional spaces, CLEAR induces a bilingual cross-lingual space that reflects the asymmetric nature of the LE relation. Such a cross-lingual LE-specialised space is illustrated in Figure 1. CLEAR is an extension of the monolingual LEAR specialisation method (Vulić and Mrkšić, 2018). The key idea of CLEAR is to leverage external lexical knowledge (i.e., information on word relations from WordNet, BabelNet, or ConceptNet) to rescale vector norms which reflect the concept hierarchy, while simultaneously pushing (i.e., "attracting") desirable word pairs closer (by vector direction) to reflect their semantic similarity in the cross-lingual LE-specialised space. Crucially, as shown later in Figure 3, CLEAR relies on a curated semantic resource only in the resource-rich source language (e.g., English): coupled with a bilingual dictionary it propagates the LE knowledge to the target (resource-poor) language and constructs a shared cross-lingual LE-specialised space. This cross-lingual LE-specialised space, depicted in Figure 1 and empirically validated in §4, is then used to reason over GR-LE in the target language, and for making cross-lingual GR-LE assertions.

Our experiments demonstrate that CLEAR is a strong benchmark on all GR-LE datasets. It can effectively transfer LE knowledge to a spectrum of target languages. What is more, through multilingual training via a resource-rich pivot language (e.g., English) CLEAR supports cross-lingual GR-LE for language pairs without any semantic resources. Finally, we report state-of-the-art scores in the *ungraded* (i.e., binary) cross-lingual LE detection for three diverse language pairs on standard evaluation sets (Upadhyay et al., 2018).

Annotation guidelines and created datasets for all languages and language pairs are available online at: `https://github.com/ivulic/xling-grle/`, and as the supplemental material. We also make available the code and CLEAR-specialised vector spaces.

2 Graded LE Evaluation Datasets

Graded lexical entailment is an asymmetric relation formulated by the intuitive question *"To what degree is X a type of Y?"*: it comprises two distinct phenomena studied in cognitive science (Hampton, 2007). First, it captures the *measure of typicality* in graded cognitive categorisation (Rosch, 1975; Medin et al., 1984): some instances of a category are more central than others (e.g., *basketball* will

often be cited as a more typical *sport* than *biathlon*). Second, it covers the measure of *vagueness* (also referred to as *graded membership*): it measures the graded applicability of a concept to different instances.[3] Despite the fact that GR-LE should not be bound to any particular surface realisation of concepts (i.e., it is not tied to a particular language), a graded LE repository has so far been created only for English: it is the HyperLex dataset of Vulić et al. (2017). Starting from the established data creation protocol for HyperLex, in this work we compile similar HyperLex datasets in three other languages and introduce novel multilingual and cross-lingual GR-LE tasks.

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 GR-LE relation. Unlike in symmetric similarity datasets (Hill et al., 2015; Gerz et al., 2016), 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 GR-LE *"To what degree..."* question to human subjects, with each pair rated by at least 10 raters: the score of 6 indicates strong 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. What is more, high inter-annotator agreement scores (see Table 3), suggest that even non-expert annotators consistently reason about the degree of LE between words.[4]

Word Pair Translation. Monolingual HyperLex datasets in three target languages: German (DE), Italian (IT), and Croatian (HR) were constructed by translating word pairs from the EN HyperLex and re-scoring 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 to ensure a wide coverage of diverse WordNet relations; **2)** we want to ensure as comparable datasets as possible across different languages in terms of semantic coverage; **3)** the approach has been extensively validated in related work on creating multilingual semantic similarity datasets (Leviant and Reichart, 2015; Camacho-Collados et al., 2017). Most importantly, the translation approach allows for the automatic construction of cross-lingual GR-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 of the target language. The translation agreement was in the range of 85%-90% across the three target languages. Translation disagreements were resolved by a third annotator who selected the correct (or better) translation following discussions with both translators. To account for polysemy, each word pair was shown along with its EN HyperLex score, helping annotators to preserve word sense during translation. We allowed for multi-word translations only if there was no appropriate single word translation (e.g., *typewriter* $\rightarrow$ *macchina da scrivere*).

Guidelines and Concept Pair Scoring. EN HyperLex annotation guidelines were translated to all three target languages (see the supplementary). The resulting 2,616 concept pairs in each language 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 native speakers.[5]

Cross-Lingual Datasets. The cross-lingual CL-HYPERLEX datasets were then constructed automatically, leveraging word pair translations and scores in three target languages. To this end, we follow the methodology of Camacho-Collados et al. (2015, 2017), used previously for creating cross-lingual semantic similarity datasets. In short, we first intersect aligned concept pairs (obtained through translation) in two languages: e.g., *father-ancestor* in English and *padre-antenato* in Italian are used

[3] Following Vulić et al. (2017), it is not clear to which extent a *washing machine* is an instance of the category *chair* despite the fact that "one can sit on washing machines".

[4] For more details on guidelines and creation of EN HyperLex we refer the reader to the original work.

[5] As opposed to (Hill et al., 2015; Gerz et al., 2016; Vulić et al., 2017), but similar to (Camacho-Collados et al., 2017; Pilehvar et al., 2018) we did not divide the dataset into smaller tranches; each annotator scored the entire target-language dataset instead. The target languages were selected based on the availability of native speakers; the total number of annotations was restricted by the annotation budget.

Monolingual Datasets			
EN	portrait	picture	5.90
DE	Idol	Person	4.0
DE	Motorrad	Fahrrad	0.25
IT	origano	cibo	3.25
HR	tenis	rekreacija	5.75
Cross-Lingual Datasets (CL-HYPERLEX)			
EN-DE	dinosaur	Kreatur	4.75
EN-IT	eye	viso	0.6
EN-HR	religija	belief	4.92
DE-IT	Medikation	trattamento	5.38
DE-HR	Form	prizma	0.0
IT-HR	aritmetica	matematika	5.5

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

	EN	DE	IT	HR
EN	2,616	3,029	3,338	3,514
DE	–	2,616	3,424	3,522
IT	–	–	2,616	3,671
HR	–	–	–	2,616

Table 2: The sizes of all monolingual (main diagonal) and cross-lingual graded LE datasets.

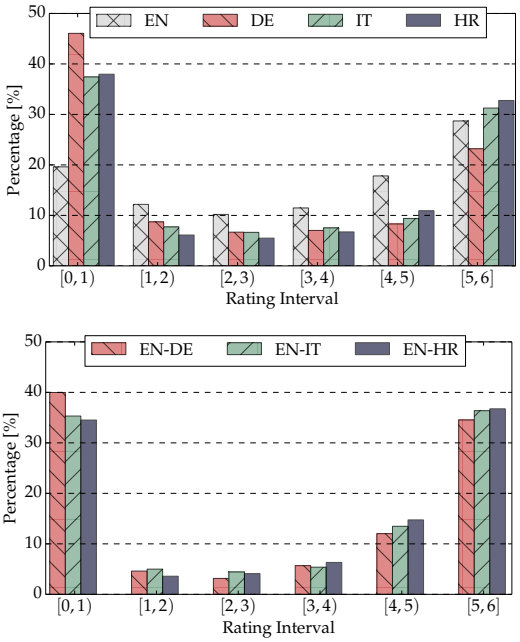


Figure 2: Rating distributions in monolingual and (a selection of) cross-lingual graded LE datasets. y axes plot percentages; the data sizes provided in Table 2.

to create cross-lingual pairs *father-antenato* and *padre-ancestor*. The GR-LE scores of cross-lingual pairs are computed as averages of corresponding monolingual scores. Finally, we retain only cross-lingual pairs for which the corresponding monolingual scores differ by $\leq$ 1.0. This heuristic (Camacho-Collados et al., 2017) mitigates the undesirable inter-language semantic shift. We refer the reader to (Camacho-Collados et al., 2015) for full (technical) description of the procedure.

Score Distributions. Table 1 displays example pairs from monolingual and cross-lingual GR-LE datasets, whereas Table 2 lists the total number of pairs for each of them. The constructed datasets are comprehensive and on a par with or larger than semantic similarity benchmarks: SimLex (Hill et al., 2015) contains 999 word pairs; multilingual and cross-lingual datasets of Camacho-Collados et al. (2017) contain $< 1,000$ pairs each. The only word similarity dataset comparable in size is SimVerb (Gerz et al., 2016) with 3,500 verb pairs. This dataset magnitude can even support supervised learning (Vulić et al., 2017; Rei et al., 2018).

We verify that all score ranges are represented by a sufficient number of concept pairs. The score distributions are shown in Figure 2. As in EN HyperLex, a large number of concept pairs is placed within the two outer sub-intervals (i.e., $[0, 1)$ and $[5, 6]$): this is an artefact of having WordNet synonyms as trivial LE pairs on the one side, whereas antonyms, no-relation, and reverse hyponymy-hypernymy pairs are found on the other side of the scoring spectrum. 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.

	EN	DE	IT	HR
Pairwise-IAA	0.854	0.741	0.736	0.840
Mean-IAA	0.864	0.803	0.809	0.882

Table 3: Inter-annotator agreement (Spearman's ρ correlation) for monolingual GR-LE datasets. IAA scores for the original EN HyperLex provided for reference.

Inter-Annotator Agreement. Following prior work on word pair dataset creation (Silberer and Lapata, 2014; Hill et al., 2015; Gerz et al., 2016; Vulić et al., 2017, *inter alia*), we report two inter-annotator agreement (IAA) measures for the three new monolingual datasets. **Pairwise-IAA** is the average pairwise Spearman's ρ correlation between any two raters. **Mean-IAA** compares the average correlation of an annotator with the average of all the other annotators: it is a human 'upper bound' for the performance of automatic systems. The IAA scores in Table 3 show that humans quantify graded

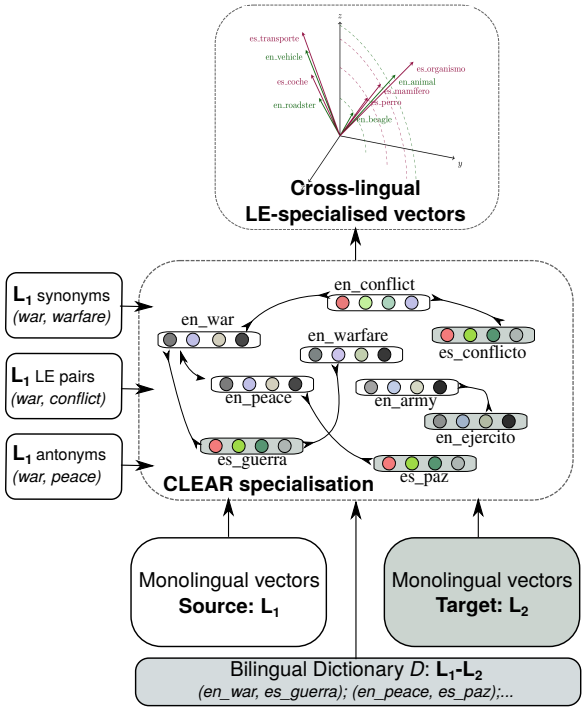


Figure 3: High-level overview (with toy examples) of the CLEAR specialisation procedure resulting in a shared cross-lingual word vector space that accentuates the LE relation between the concepts.

LE consistently across languages.[6] High Mean-IAA scores are challenging upper bounds that justify our automatic construction of CL-HYPERLEX.

We further validate CL-HYPERLEX by comparing automatically induced scores with human judgements. For each EN-{DE,IT,HR} dataset we let two annotators fluent in both languages judge 333 randomly sampled pairs. We report high average Spearman's ρ correlation between automatically induced scores and human judgements: 0.896 (EN-DE), 0.909 (EN-IT), and 0.905 (EN-HR).

3 Methodology

In order to provide benchmarking graded LE scores on new monolingual and cross-lingual evaluation sets, we now introduce a novel method that can capture GR-LE cross-lingually. CLEAR (**C**ross-**L**ingual **L**exical **E**ntailment **A**ttract-**R**epel) is a cross-lingual extension of the monolingual LEAR specialisation method (Vulić and Mrkšić, 2018), a state-of-the-art vector space fine-tuning method which specialises any input distributional vector space to accentuate the asymmetric LE relation in the transformed space. We show that, coupled with a bilingual dictionary, CLEAR can learn vector rearrangements that reflect lexical entailment also in the target language for which no external lexical knowledge concerning the LE relation is available, and it can also quantify the degree of cross-lingual LE. The core idea is to simultaneously capture the hierarchy of concepts (through vector norms) and their similarity (through their cosine distance), irrespective of the actual language (see Figure 1).

CLEAR Specialisation. A high-level overview of the CLEAR specialisation method is provided in Figure 3. The input to the method is as follows: **1)** two independently trained monolingual word vector spaces in two languages L_1 and L_2; **2)** sets of external lexical constraints in the resource-rich language L_1 (e.g., English) extracted from an external lexical resource such as WordNet (Fellbaum, 1998) or BabelNet (Ehrmann et al., 2014); and **3)** a bilingual L_1-L_2 dictionary D. The goal is to fine-tune input word vectors in both languages using the L_1 lexical constraints and the dictionary D, and obtain a shared cross-lingual space specialised for LE.

CLEAR uses a set of external linguistic constraints $C = S \cup A \cup Le$ in language L_1 for fine-tuning. The set comprises synonymy pairs S such as *(clever, smart)*, antonymy pairs A such as *(war, peace)*, and lexical entailment (i.e., hyponymy-hypernymy) pairs Le such as *(dog, animal)*. For the Le pairs, the word order is important: we assume that the left word is always the hyponym. Further, we treat pairs from the dictionary D such as *(war, guerra)* as another distinct set of (cross-lingual) synonymy pairs. The D pairs are L_1-L_2 pairs, while all the remaining word pairs are L_1 pairs: this creates a true cross-lingual transfer setup.

Similar to LEAR and the ATTRACT-REPEL model for symmetric similarity specialisation (Mrkšić et al., 2017), CLEAR defines two types of *symmetric* objectives for the L_1 pairs: 1) the ATTRACT (*Att*) objective aims to bring closer together in the vector space words that are semantically similar (i.e., synonyms and hyponym-hypernym pairs); 2) the REPEL (*Rep*) objective pushes apart vectors of dissimilar words (i.e., antonyms). We denote as $\mathcal{B} = \{(\mathbf{x}_l^{(k)}, \mathbf{x}_r^{(k)})\}_{k=1}^{K}$ the set of K word vector pairs for which the *Att* or *Rep* score is to be computed: we refer to these pairs as the *positive examples*. The set of corresponding negative examples T is created by coupling each positive AT-

[6] Similarity benchmarks report much lower Pairwise-IAA scores: 0.61 on SimVerb-3500 (Gerz et al., 2016; Pilehvar et al., 2018), and 0.67 on SimLex-999 (Hill et al., 2015) and on WordSim-353 (Finkelstein et al., 2002)

TRACT example $(\mathbf{x}_l, \mathbf{x}_r)$ with a negative example pair $(\mathbf{t}_l, \mathbf{t}_r)$, where $\mathbf{t}_l$ is the vector closest (within the current batch in terms of cosine similarity) to $\mathbf{x}_l$, and $\mathbf{t}_r$ the vector closest to $\mathbf{x}_r$. The *Att* objective $Att(\mathcal{B}_{Att}, T_{Att})$ for a batch of ATTRACT constraints $\mathcal{B}_{Att}$ is then formulated as the max-margin learning problem as follows:

$$\sum_{k=1}^{K} [\, \tau\left(\delta_{att} + \cos\left(\mathbf{x}_l^{(k)}, \mathbf{t}_l^{(k)}\right) - \cos\left(\mathbf{x}_l^{(k)}, \mathbf{x}_r^{(k)}\right)\right) + \tau\left(\delta_{att} + \cos\left(\mathbf{x}_r^{(k)}, \mathbf{t}_r^{(k)}\right) - \cos\left(\mathbf{x}_l^{(k)}, \mathbf{x}_r^{(k)}\right)\right)]. \quad (1)$$

$\tau(x) = max(0, x)$ is the ramp function and δ_{att} is the similarity margin imposed between the negative and positive vector pairs. The *Rep* objective is designed in a similar fashion: for each positive REPEL example, the negative example $(\mathbf{t}_l, \mathbf{t}_r)$ couples the vector $\mathbf{t}_l$ that is most distant from $\mathbf{x}_l$ (cosine similarity in the current batch) and $\mathbf{t}_r$, most distant from $\mathbf{x}_r$. The goal of the *Rep* objective $Rep(\mathcal{B}_{Rep}, T_{Rep})$ for a batch of REPEL word pairs $\mathcal{B}_{Rep}$ and the corresponding negative examples T_{Rep} is then to push REPEL pairs away from each other by the "repel" margin δ_{rep}. The exact formulation is analogous to the *Att* objective, and not shown for brevity.

Crucially, similar to LEAR, CLEAR forces specialised vectors to reflect the asymmetry of the LE relation with an asymmetric distance-based objective. Starting from the Le (hyponymy-hypernymy) pairs, the goal is to rearrange vectors of words in these pairs, that is, to preserve the cosine distances in the specialised space while steering vectors of more general concepts to take larger norms, as shown in Figure 1 and 3. We adopt the best-performing asymmetric objective from Vulić and Mrkšić (2018) and use it with L_1 Le word pairs:

$$LE(\mathcal{B}_{Le}) = \sum_{k=1}^{K} \frac{\|\mathbf{x}_l^{(k)}\| - \|\mathbf{x}_r^{(k)}\|}{\|\mathbf{x}_l^{(k)}\| + \|\mathbf{x}_r^{(k)}\|}. \quad (2)$$

The objectives described so far cover S, A, and Le word pairs. The translation pairs from the dictionary D are also "attracted" to each other, but using a different objective. We define the $Att_D(\mathcal{B}_D)$ objective on a batch of translation pairs $\mathcal{B}_D$ as the simple ℓ_2-distance between two words in each pair:

$$Att_D(\mathcal{B}_D) = \lambda_D \sum_{k=1}^{K} \|\mathbf{x}_l^{(k)} - \mathbf{x}_r^{(k)}\|. \quad (3)$$

$\mathbf{x}_l^{(k)}$ is the vector of an L_1 word from the source language vector space and $\mathbf{x}_r^{(k)}$ the vector of its L_2 translation from the target language space. λ_D is the cross-lingual regularisation factor. The rationale behind this design is as follows: in order to rearrange word vectors of *both* languages as shown in Figure 1, we have to allow for the adjustment of vector norms also for L_2 word vectors. The previous Att objective from Eq. (1) relies on the cosine similarity and captures only the vector direction.

Finally, CLEAR defines a regularisation term for all word pairs in the sets S, A, Le, and D in order to preserve the useful semantic information from the original distributional spaces. Let $V(\mathcal{B})$ denote the set of distinct words in a constraint batch $\mathcal{B}$; the regularisation term is then: $Reg(\mathcal{B}) = \lambda_{reg} \sum_{\mathbf{x} \in V(\mathcal{B})} \|\mathbf{y} - \mathbf{x}\|_2$, where $\mathbf{y}$ is the CLEAR-transformed vector of any distributional vector $\mathbf{x}$, and λ_{reg} is the regularisation factor. The full CLEAR objective is then defined as follows:

$$\begin{aligned} J = & \; Att(\mathcal{B}_S, T_S) + Rep(\mathcal{B}_A, T_A) \\ & + Att(\mathcal{B}_{Le}, T_{Le}) + LE(\mathcal{B}_{Le}) \\ & + Att_D(\mathcal{B}_D) + Reg(\mathcal{B}_S, \mathcal{B}_A, \mathcal{B}_{Le}, \mathcal{B}_D) \end{aligned} \quad (4)$$

This joint objective rearranges vectors from both input monolingual vector spaces (see Figure 3) and enables the transfer of LE signal from the resource-rich language L_1 to the target language (i.e., CLEAR does not rely on any explicit LE knowledge in L_2).

Asymmetric LE Distance. Monolingual and cross-lingual LE strength can be inferred directly from the CLEAR-specialised cross-lingual space. It is done by a distance function that reflects both the cosine distance between the vectors (semantic similarity) as well as the asymmetric difference between the vectors' norms (Vulić and Mrkšić, 2018):

$$I_{LE}(\mathbf{x}, \mathbf{y}) = dcos(\mathbf{x}, \mathbf{y}) + \frac{\|\mathbf{x}\| - \|\mathbf{y}\|}{\|\mathbf{x}\| + \|\mathbf{y}\|} \quad (5)$$

$\mathbf{x}$ and $\mathbf{y}$ are vectors of *any* two words x and y in the cross-lingual space. For less expressive ungraded LE detection tasks I_{LE} distances are trivially transformed into binary LE predictions using a binarisation threshold t: if $I_{LE}(\mathbf{x}, \mathbf{y}) < t$, we predict that the LE relation holds between words x and y. CLEAR-specialized vectors of general concepts obtain larger norms than vectors of specific concepts. Strong LE pairs should display both small cosine distances and negative norm differences.

4 Results and Discussion

We run experiments with representative baseline models and CLEAR-specialised vectors on new

monolingual and cross-lingual graded LE datasets, as well as on established ungraded cross-lingual LE detection datasets (Vyas and Carpuat, 2016; Upadhyay et al., 2018). The goal of reported experiments is twofold: besides providing baseline scores on new evaluation sets, we also analyse the usefulness of cross-lingual graded LE specialisation performed by CLEAR, and analyse its performance in comparison with distributional word vectors and non-specialised cross-lingual word embeddings.

4.1 Experimental Setup

Distributional Vectors. Graded LE is evaluated on EN, DE, IT, and HR (see §2); we also evaluate CLEAR on ungraded cross-lingual LE (Upadhyay et al., 2018) for the following language pairs: EN-FR (French); EN-RU (Russian); EN-AR (Arabic).

All results are reported with English Skip-Gram with Negative Sampling (SGNS-BOW2) vectors (Mikolov et al., 2013) trained by Levy and Goldberg (2014) on the Polyglot Wikipedia (Al-Rfou et al., 2013) with bag-of-words context (window size of 2).[7] Input vectors for other languages come from various sources: AR vectors are `fastText` vectors trained on the Common Crawl data by Grave et al. (2018). RU vectors are obtained by Kutuzov and Andreev (2015). FR, IT, DE, and HR word vectors are large SGNS vectors trained on the standard frWaC, itWaC, and deWaC corpora (Baroni et al., 2009), and the hrWaC corpus (Ljubešić and Klubička, 2014), also used in prior work (Vulić et al., 2017). All word vectors are 300-dim.[8]

Linguistic Constraints and Dictionaries. We use the same set of monolingual constraints as LEAR (Vulić and Mrkšić, 2018): synonymy and antonymy constraints from (Zhang et al., 2014; Ono et al., 2015) are extracted from WordNet and Roget's Thesaurus (Kipfer, 2009). As in other work on LE specialisation (Nguyen et al., 2017; Nickel and Kiela, 2017), asymmetric LE constraints are extracted from WordNet, and we collect both direct and indirect LE pairs (i.e., *(beagle, dog)*, *(dog, animal)*, and *(beagle, animal)* are in the *Le* set) In total, we work with 1,023,082 pairs of synonyms, 380,873 pairs of antonyms, and 1,545,630 LE pairs.

Bilingual dictionaries are derived from PanLex (Kamholz et al., 2014), which was used in prior work on cross-lingual word embeddings (Duong et al., 2016; Adams et al., 2017; Vulić et al., 2017). PanLex currently spans around 1,300 language varieties with over 12M expressions: it offers support also to low-resource transfer settings.[9]

Training Setup. CLEAR hyperparameters are adopted from the original Attract-Repel work (Mrkšić et al., 2017): $\delta_{att} = 0.6$, $\delta_{rep} = 0.0$, $\lambda_{reg} = \lambda_D = 10^{-9}$. All batches are of size 128 (see Eq. (4)), and the model is trained for 5 epochs with Adagrad (Duchi et al., 2011).

Baseline Models. In monolingual evaluation, we compare CLEAR to original non-specialised distributional vectors in each language. Another instructive baseline is the TRANS baseline which uses exactly the same amount of information as CLEAR. Instead of performing joint CLEAR specialisation as described in §3, TRANS is a two-step process that: 1) runs the monolingual LEAR specialisation of the English distributional space, and then 2) translates all test examples in the target language to English relying on the bilingual dictionary D.[10] All LE reasoning is then conducted monolingually in English.

The TRANS baseline is also used in cross-lingual graded LE evaluation. For cross-lingual datasets without English (e.g., DE-IT), we again translate all words to English and use the English specialised space for graded LE assertions. In addition, for each language pair we also report results of two state-of-the-art cross-lingual word embedding models (Smith et al., 2017; Artetxe et al., 2018), showing the better scoring one in each run (XEMB).

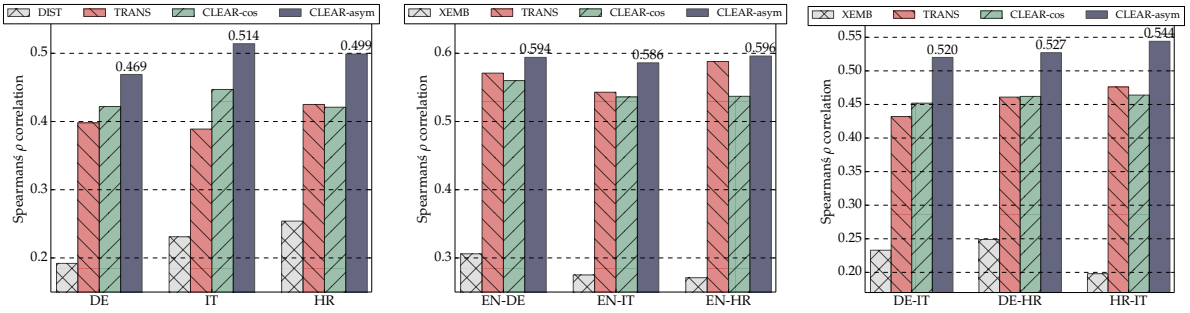
For ungraded LE evaluation, in addition to TRANS, we compare CLEAR to two best-performing baselines from (Upadhyay et al., 2018): they couple two methods for inducing syntactic cross-lingual vectors: 1) BI-SPARSE (Vyas and Carpuat, 2016) and 2) CL-DEP (Vulić, 2017) with an LE scorer based on the distributional inclusion hypothesis (Geffet and Dagan, 2005). For more details we refer the reader to (Upadhyay et al., 2018).

[7] The proposed CLEAR method is by design agnostic of input distributional vectors and its main purpose is to support fine-tuning of a wide spectrum of input vectors. We have experimented with other standard distributional spaces in English such as `fastText` (Bojanowski et al., 2017; Grave et al., 2018), type-based ELMo embeddings (Peters et al., 2018), Context2Vec (Melamud et al., 2016) and Glove (Pennington et al., 2014), but the obtained results follow similar trends. We do not report these results for brevity.

[8] Vectors of multi-word expressions in CL-HYPERLEX are obtained by averaging over their constituent words' vectors.

[9] The translations in PanLex were derived from various sources (e.g., glossaries, dictionaries, automatic inference). This results in high-coverage but noisy lexicons.

[10] In cases where one word has more than one EN translation, we randomly sample a single translation from D.



(a) Monolingual (b) CL-HYPERLEX: EN-L_2 (c) CL-HYPERLEX: Other

Figure 4: Summary of monolingual and cross-lingual graded LE results (Spearman's ρ correlation scores). **(a)** Monolingual evaluation on target languages; **(b)** Cross-lingual evaluation with EN included in each pair; **(c)** Cross-lingual evaluation: the scores are obtained via multilingual training of a joint EN-DE-IT-HR CLEAR model.

4.2 Results and Discussion

Graded LE Evaluation. First, we evaluate the transfer capability of CLEAR: we make graded LE assertions monolingually in each target language without seeing a single hyponymy-hypernymy pair in the target, and evaluate the method on newly constructed monolingual HyperLex datasets. The results (Spearman's ρ) are summarised in Figure 4a. They suggest that the CLEAR transfer is a viable strategy for LE-specialising target language vector spaces. Non-specialised input distributional vectors are not suitable for capturing graded LE. Importantly, CLEAR outperforms the direct translation approach (TRANS). Furthermore, the comparison between two CLEAR configurations reveals that the asymmetric distance (see Eq. (5)) is indeed crucial for improved performance: we observe consistent gains with the CLEAR-asym model, which uses full I_{LE} from Eq. (5) for inference, over CLEAR-cos, which relies only on the symmetric cosine distance $dcos$, without leveraging vector norms.

The results on three EN-{DE, IT, HR} cross-lingual graded LE datasets are provided in Figure 4b. They largely follow the patterns already established in the monolingual graded LE task: non-specialised cross-lingual word vectors cannot match performance of other models, and CLEAR-asym is the strongest model across the board.

To verify that CLEAR is not to tied to any specific dictionary, we have also experimented with cross-lingual BabelNet synsets (Ehrmann et al., 2014), and combined BabelNet+PanLex dictionaries leading to very similar trends in results, with PanLex showing a slight edge over BabelNet. Furthermore, we leave experiments with dictionaries induced by unsupervised and weakly supervised cross-lingual word embeddings (Conneau et al., 2018a; Artetxe et al., 2018; Glavaš et al., 2019) for future work.

We also provide results on other cross-lingual datasets relying on multilingual training: we fix EN as the single source language and propagate LE information to multiple target languages. To this end, we train a four-lingual EN-DE-IT-HR model. The main finding from Figure 4c is that multilingual training can effectively LE-specialise target language vector spaces and enable reasoning over the cross-lingual graded LE relation even in settings with limited or no target lexico-semantic resources.

Finally, additional multilingual knowledge introduced through dictionaries D and distributional spaces of target languages is also beneficial for monolingual GR-LE in the resource-rich language. Previous best results on the EN HyperLex were 0.686 on the entire dataset and 0.703 on its noun portion (Vulić and Mrkšić, 2018). All bilingual EN-L_2 CLEAR models surpass these scores: e.g., the EN-IT model scores 0.691 on the entire dataset (0.712 on noun pairs). The best result on EN HyperLex is reported with the four-lingual CLEAR EN-DE-IT-HR model: 0.701 (0.719 on nouns).

Ungraded Cross-Lingual LE Evaluation. We further demonstrate the effectiveness of CLEAR on ungraded cross-lingual LE benchmarks from Upadhyay et al. (2018). The models are evaluated on two types of test sets: HYPO – where LE pairs need to be distinguished from inverse LE (i.e., hypernym-hyponym) pairs and COHYP in which LE pairs are to be differentiated from cohyponyms. Each test set has a corresponding train portion, which we use to tune the binarisation threshold t for I_{LE} scores.

The ungraded cross-lingual LE performance of CLEAR for three diverse language pairs (EN-FR, EN-RU, EN-AR) is shown in Table 4. The results prove CLEAR's robustness for cross-lingual LE modeling:

	Model	EN-FR	EN-RU	EN-AR
HYPO	CL-DEP	0.538	0.602	0.567
	BI-SPARSE	0.566	0.590	0.526
	TRANS	0.766	0.764	0.690
	CLEAR	**0.821**	**0.791**	**0.783**
COHYP	CL-DEP	0.610	0.562	0.631
	BI-SPARSE	0.667	0.636	0.668
	TRANS	0.759	0.751	0.696
	CLEAR	**0.885**	**0.871**	**0.814**

Table 4: Cross-lingual ungraded LE detection accuracy scores on cross-lingual HYPO and COHYP evaluation sets from Upadhyay et al. (2018).

it substantially outperforms (by 22% on average) the current state-of-the-art models BI-SPARSE and CL-DEP (Upadhyay et al., 2018) in both HYPO and COHYP tasks, and for all language pairs. CLEAR again shows that it can LE-specialise target vectors without any target-language LE knowledge. It displays highest performance for EN-FR, but the drop in performance for EN-RU and EN-AR, is not large (especially for the HYPO setting).

Extending CLEAR. As the main goal of this work is to validate the cross-lingual transfer potential and wide portability of the CLEAR model, we do not leverage any target language constraints. However, note that further improvements are expected by explicitly injecting symmetric and asymmetric linguistic constraints in the target language, if these are available, e.g., from BabelNet or multilingual WordNet (Bond and Foster, 2013).

We also stress that the CLEAR method inherits the main "retrofitting" property of the underlying monolingual LEAR method: it updates (i.e., LE-specialises) only the vectors of words which are observed in the sets of external linguistic constraints. We believe that further improvements of the CLEAR transfer method can be achieved by LE-specialising the full distributional spaces through recently proposed post-specialisation methods which learn a global specialisation function (Ponti et al., 2018; Kamath et al., 2019; Glavaš and Vulić, 2018; Glavaš and Vulić, 2019).

5 Conclusion and Future Work

We have proposed a novel graded cross-lingual lexical entailment (LE) task, introducing new monolingual and cross-lingual graded LE datasets that hold promise to support future research on this topic. We have then proposed a transfer-based method that can reason over graded LE across languages. We have demonstrated its robustness and usefulness for graded and ungraded LE in monolingual and cross-lingual settings. In the future, we will work on cross-lingual extensions of monolingual hyperbolic embedding models (Nickel and Kiela, 2017; Ganea et al., 2018). We will also experiment with other sources of bilingual information (e.g., cross-lingual word embeddings) and port the transfer approach to more language pairs, with a particular focus on resource-poor languages.

Evaluation data for multilingual and cross-lingual graded LE is available online at: `github.com/ivulic/xling-grle/`.

Acknowledgments

We thank our annotators for helping us create multilingual and cross-lingual HyperLex resources, and the three anonymous reviewers for their helpful suggestions. Goran Glavaš is supported by the Baden-Württemberg Stiftung's Eliteprogramm grant AGREE (*"Algebraic Reasoning over Events from Text and External Knowledge"*).

References

Oliver Adams, Adam Makarucha, Graham Neubig, Steven Bird, and Trevor Cohn. 2017. Cross-lingual word embeddings for low-resource language modeling. In *Proceedings of EACL*, pages 937–947.

Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2013. Polyglot: Distributed word representations for multilingual NLP. In *Proceedings of CoNLL*, pages 183–192.

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.

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The WaCky wide web: A collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, 43(3):209–226.

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.

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

Francis Bond and Ryan Foster. 2013. Linking and extending an open multilingual Wordnet. In *Proceedings of ACL*, pages 1352–1362.

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.

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.

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

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018a. Word translation without parallel data. In *Proceedings of ICLR (Conference Track)*.

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

John C. Duchi, Elad Hazan, and Yoram Singer. 2011. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12:2121–2159.

Long Duong, Hiroshi Kanayama, Tengfei Ma, Steven Bird, and Trevor Cohn. 2016. Learning crosslingual word embeddings without bilingual corpora. In *Proceedings of EMNLP*, pages 1285–1295.

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.

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

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. 2002. Placing search in context: The concept revisited. *ACM Transactions on Information Systems*, 20(1):116–131.

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.

Octavian-Eugen Ganea, Gary Bécigneul, and Thomas Hofmann. 2018. Hyperbolic entailment cones for learning hierarchical embeddings. In *Proceedings of ICML*, pages 1632–1641.

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, 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*.

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

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

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of LREC*, pages 3483–3487.

James A. Hampton. 2007. Typicality, graded membership, and vagueness. *Cognitive Science*, 31(3):355–384.

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*.

David Kamholz, Jonathan Pool, and Susan M. Colowick. 2014. PanLex: Building a resource for panlingual lexical translation. In *Proceedings of LREC*, pages 3145–3150.

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.

Barbara Ann Kipfer. 2009. *Roget's 21st Century Thesaurus (3rd Edition)*. Philip Lief Group.

Lili Kotlerman, Ido Dagan, Idan Szpektor, and Maayan Zhitomirsky-Geffet. 2010. Directional distributional similarity for lexical inference. *Natural Language Engineering*, 16(4):359–389.

Andrey Kutuzov and Igor Andreev. 2015. Texts in, meaning out: neural language models in semantic similarity task for Russian. In *Proceedings of DIALOG*.

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 and Yoav Goldberg. 2014. Dependency-based word embeddings. In *Proceedings of ACL*, pages 302–308.

Nikola Ljubešić and Filip Klubička. 2014. {bs,hr,sr}WaC – Web corpora of Bosnian, Croatian and Serbian. In *Proceedings of the 9th Web as Corpus Workshop*, pages 29–35, Gothenburg, Sweden. Association for Computational Linguistics.

Douglas L. Medin, Mark W. Altom, and Timothy D. Murphy. 1984. Given versus induced category representations: Use of prototype and exemplar information in classification. *Journal of Experimental Psychology*, 10(3):333–352.

Oren Melamud, David McClosky, Siddharth Patwardhan, and Mohit Bansal. 2016. The role of context types and dimensionality in learning word embeddings. In *Proceedings of NAACL-HLT*, pages 1030–1040.

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

Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of NIPS*, pages 3111–3119.

Nikola Mrkšić, Diarmuid Ó Séaghdha, Tsung-Hsien Wen, Blaise Thomson, and Steve Young. 2017. Neural belief tracker: Data-driven dialogue state tracking. In *Proceedings of ACL*, pages 1777–1788.

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.

Matteo Negri, Alessandro Marchetti, Yashar Mehdad, Luisa Bentivogli, and Danilo Giampiccolo. 2012. Semeval-2012 task 8: Cross-lingual textual entailment for content synchronization. In *Proceedings of SEMEVAL*, pages 399–407.

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 NIPS*, pages 6341–6350.

Masataka Ono, Makoto Miwa, and Yutaka Sasaki. 2015. Word Embedding-based Antonym Detection using Thesauri and Distributional Information. In *Proceedings of NAACL*, pages 984–989.

Sebastian Padó, Michel Galley, Dan Jurafsky, and Christopher D. Manning. 2009. Robust machine translation evaluation with entailment features. In *Proceedings of ACL*, pages 297–305.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of EMNLP*, pages 1532–1543.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of NAACL-HLT*, pages 2227–2237.

Mohammad Taher Pilehvar, Dimitri Kartsaklis, Victor Prokhorov, and Nigel Collier. 2018. Card-660: Cambridge rare word dataset - a reliable benchmark for infrequent word representation models. In *Proceedings of EMNLP*, pages 1391–1401.

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.

Marek Rei, Daniela Gerz, and Ivan Vulić. 2018. Scoring lexical entailment with a supervised directional similarity network. In *Proceedings of ACL*, pages 638–643.

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. 1973. Natural categories. *Cognitive Psychology*, 4(3):328–350.

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.

Carina Silberer and Mirella Lapata. 2014. Learning grounded meaning representations with autoencoders. In *Proceedings of ACL*, pages 721–732.

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 (Conference Track)*.

Zhiyi Song, Ann Bies, Justin Mott, Xuansong Li, Stephanie M. Strassel, and Christopher Caruso. 2018. Cross-document, cross-language event coreference annotation using event hoppers. In *Proceedings of LREC*, pages 3535–3540.

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

Peter D. Turney and Saif M. Mohammad. 2015. Experiments with three approaches to recognizing lexical entailment. *Natural Language Engineering*, 21(3):437–476.

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. 2017. Cross-lingual syntactically informed distributed word representations. In *Proceedings of EACL*, pages 408–414.

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

Ivan Vulić, Nikola Mrkšić, and Anna Korhonen. 2017. Cross-lingual induction and transfer of verb classes based on word vector space specialisation. In *Proceedings of EMNLP*, pages 2536–2548.

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

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.

Jingwei Zhang, Jeremy Salwen, Michael Glass, and Alfio Gliozzo. 2014. Word semantic representations using bayesian probabilistic tensor factorization. In *Proceedings of EMNLP*, pages 1522–1531.