Building MUSCLE, a Dataset for MUltilingual Semantic Classification of Links between Entities

Lucia Pitarch, Carlos Bobed, David Abián, Jorge Gracia, Jorge Bernad
University of Zaragoza
{lpitarch, cbobed, abian, jbernad, jogracia}@unizar.es

Abstract

In this paper we introduce MUSCLE, a dataset for MUltilingual lexico-Semantic Classification of Links between Entities. The MUSCLE dataset was designed to train and evaluate Lexical Relation Classification (LRC) systems, and contains 27K pairs of universal concepts selected from Wikidata, a large and highly multilingual factual Knowledge Graph (KG). Each pair of concepts includes lexical forms in 25 languages and is labeled with up to five possible lexico-semantic relations between the concepts: hypernymy, hyponymy, meronymy, holonymy, and antonymy. Inspired by Semantic Map theory, the dataset bridges lexical and conceptual semantics, is more challenging and robust than previous datasets for LRC, avoids lexical memorization, is domain-balanced across entities, and enables enrichment and hierarchical information retrieval. To establish baseline results for further research, we also evaluate the dataset in LRC under a minimal prompting setting, providing a comparison to datasets such as K&H+N, BLESS, CogALexV, EVALution and ROOT9 for English, and CogALexVI in its multilingual setting.
Keywords: Lexical Multilingual Dataset, Lexical Relation Classification, Semantic Maps, Conceptual Grounding

1. Introduction

Lexico-semantic relations are abstract semantic and lexical construals negotiated between speakers that help structure how we group meanings and approach the world (Murphy, 2003). Such relations are the basis of fundamental linguistic resources such as WordNets (Miller, 1995): they help organize the lexicon and how we linguistically encode the world. The main lexico-semantic relations are 1) **synonymy**, which links similar meanings (and is used to create WordNet nodes called synsets); 2) **antonymy**, which links contrasting meanings; 3) **hyponymy/hypernymy**, which establishes hierarchy; and 4) **meronymy/holonymy**, which represents part-whole relations.

The automatic classification of such lexico-semantic relations, known as computational Lexical Relation Classification (LRC), has received continuous attention due to its impact on several NLP tasks (Necsulescu et al., 2015) and key lexical resources such as WordNets. Methods used have ranged from path-based methods (Hearst, 1992) to distributional methods using Pre-Trained Language Models (PTLMs), the current SoTA (Pitarch Ballesteros et al., 2023).

Despite this progress, LRC is still challenging because lexico-semantic relations are neither fully universal nor static, but rather malleable and adaptable to particular contexts, and not even humans sometimes agree on how to annotate them (Stevenson and Merlo, 2022). For example, *boy* and *man* can be antonyms when the 'age' feature is relevant, but synonyms when it is not. Furthermore, while PTLMs' underlying semantics are distributional, they lack conceptual grounding, making them too reliant on the co-occurrence of lexical entries in texts. And, while this does not mean they do not have deeper conceptual semantic knowledge, it does not ensure it either. To deal with this limitation, in this work, we adopt the notion of Semantic Maps as defined by François (2008), which, starting from lexicalizations, provide a method for delimiting concepts by finding their commonalities in different languages to uncover semantic patterns universal to human perception and cognition. This makes it possible to surpass distributional semantics (available in Language Models), establishing some other referent to the target meaning by finding what remains common to all multilingual verbalizations. With this approach we intend to go a step further towards Conceptual Grounding (Silberer and Lapata, 2012) where they claim meaning comes not only from word's distributional properties, but also from the personal experience and interaction with the world of the people using them. Defining concepts by grouping lexical forms in different languages was already implemented by datasets such as CLICS3 (Rzymski et al., 2020). Yet, to the best of our knowledge, Semantic Maps have not been applied to LRC, a gap we fill with our dataset.

The main contribution of this paper is the development of MUSCLE[1], a Dataset for MUltilingual Semantic Classification of Links between Entities, aimed at bridging the lexico-semantic gap in Language Models by applying Semantic Maps. MUSCLE has a number of added benefits in contrast to previous datasets for LRC, namely:

1. **Multilinguality**: Research regarding LRC has mainly focused on developing monolingual studies and datasets, particularly for English (Baroni and Lenci, 2011a; Necsulescu et al., 2015; Santus et al., 2016b, 2015, 2016a). Fewer proposals, such as CogALexVI (Xiang et al., 2020) and its refinement CogALex 2.0 (Lang et al., 2021), have taken a multilingual approach using English, German, Mandarin Chinese, and Italian in test data. We further extend the multilinguality up to 25 languages in our current proposal.
2. **Avoidance of lexical memorization:** When

[1]The MUSCLE dataset and the code to reproduce the experiments are available at https://github.com/sid-unizar/MUSCLE.

LREC-COLING 2024, pages 2580–2594

duplicated tokens appear in the training and test sets, the model used to solve the LRC task might be prone to memorizing that such repeated tokens should always be assigned to the same relation (Levy et al., 2015). Providing a dataset that allows conceptual grounding in fine-tuning mitigates this issue and enables checking it.

3. **Semantic domain analysis and avoidance of semantic memorization:** Similarly to lexical memorization, but at the semantic level, words[2] belonging to the same conceptual domain should not be always assigned to the same relation. Keeping the concepts underlying the lexical forms in different languages within the dataset makes it possible to check for semantic memorization issues and to perform deeper semantic analyses, including hierarchical information analyses and community detection and distribution analyses.

The concepts, relations, and lexical entries from the MUSCLE dataset were selected from Wikidata (Vrandečić and Krötzsch, 2014), a large and highly multilingual factual Knowledge Graph (KG). This is a novel data source for LRC, in contrast to previous LRC datasets which were all created using the same data sources, mainly WordNets (Miller, 1995) and ConceptNet (Liu and Singh, 2004). Using Wikidata, we aim at enhancing generalizability and bias identification. MUSCLE consists of over 27K concepts with lexical forms in 25 languages.

The paper is structured as follows. First, we present the background and related work in Section 2. Then, we define the requirements and detail the design of the MUSCLE dataset in Section 3, analysing its semantic distribution in Section 4. We describe the different configurations of the dataset in Section 5 and evaluate them in Section 6. Finally, we discuss the outcomes and future work in Section 7.

2. Background and Related Work

Similar sets of words tend to be used in similar contexts. This enables Distributional Semantic Models (DSM) to embed words in continuous vector spaces and use them to predict the most likely word for a given context (Boleda, 2019). However, just describing words by the co-occurrence of lexical forms is not enough: conceptual grounding of such words is also needed to fully describe their meaning.

Although LRC benefits from both lexical forms and conceptual groundings of words (Storjohann, 2015), previously used datasets (e.g. BLESS (Baroni and Lenci, 2011a), EVALution (Santus et al., 2015), CogALex-V (Santus et al., 2016a), ROOT9 (Santus et al., 2016b), CogALex-VI (Xiang et al., 2020), CogALex 2.0 (Lang et al., 2021)) have de facto omitted this information, providing just lexical entries (i.e., with no additional conceptual grounding) to evaluate Distributional

[2]We are using 'word' in a non restrictive way along the paper, including compound words and multi-word expressions.

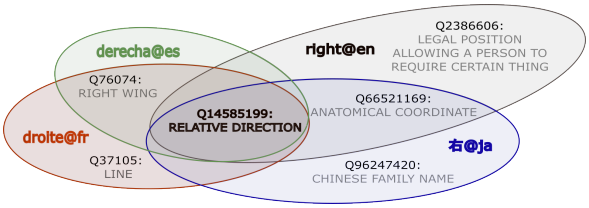

Figure 1: Semantic map for *RIGHT*

Models on LRC. To fill this gap, our proposal takes inspiration from Semantic Map theory (Haspelmath, 2003), and particularly from its application to lexical items by (François, 2008), to propose a dataset that can take into account both lexical and semantic levels for LRC.

In the structuralist approach to semantics, meaning is defined through negation. That is, meaning is defined by explicitly defining what it is not, and consequently implicitly stating what the concept does cover. François (2008) takes this idea and uses cross-lingual comparison to delimit what a concept is by finding the semantic commonalities between words that to some extent verbalize[3] the same concept.

In previous datasets for LRC, only relations between lexical forms were provided (e.g., *right* and *left* as antonyms). Although these lexical forms were retrieved from sources such as WordNets (Miller, 1995) and ConceptNet (Liu and Singh, 2004), which assign a language-independent identifier to the target sense of a given word (e.g., bn:00067808n for the "relative direction" sense of the word 'right'), these identifiers were not included in the datasets. Nor would it be sufficient to feed a Language Model (LM) with these identifiers, since the LM would inevitably learn a new, separate embedding to encode each of them. Furthermore, while WordNets are based on synonym clusters, and by doing so they are intrinsically additive, our definition of concepts is quite the opposite as we use cross-lingual comparison not to cluster, but rather, distill the core attributes shared among all 25 languages.

The derived research question for this is: How can we conceptually ground the model just through polysemous words such as 'right'? The MUSCLE dataset main building blocks are concepts (e.g., 'right' as a "relative direction", Wikidata's Q14565199), including lexical forms for each concept in 25 languages (e.g., 'right', 'derecha', 'droite', ...), which sets MUSCLE apart from previous datasets for LRC. By fine tuning a model with several lexical forms at once for the same concept, the model has the opportunity to resolve the ambiguities of each particular form and learn the concept based on shared semantics. Figure 1 illustrates how comparison of WikiData concepts and labels can be used to decouple universal properties underlying concepts in MUSCLE, following Semantic Maps theory.

Previous datasets for LRC were all created using the same data sources, mainly WordNets (Miller, 1995),

[3]In linguistics (François, 2008), *lexification* is sometimes used to name what we call here *verbalization*, *lexical form*, or *label* of a term in KG terminology.

ConceptNet (Liu and Singh, 2004), and McRae's Norms (McRae et al., 2005). Such LRC datasets are typically modified versions of previous ones, which increases the possibility of unseen bias propagating from one dataset to another. For instance, EVALution was built on BLESS and expanded it by adding new relations and domain data, then subsampled and stemmed in CogALex-V, and again expanded in ROOT9. This was subsequently enriched with other languages in CogALex-VI and partially curated in CogALex 2.0. By contributing with a new dataset based on a different data source, we help to check the generalizability of LRC models.

MUSCLE, a large multilingual dataset that provides lexico-semantic relations in 25 languages, represents a step towards addressing these limitations. In the creation of the dataset, we checked for many relevant possible issues, including those mentioned in previous LRC research as prototypicality (Santus et al., 2015) and lexical memorization (Levy et al., 2015), as well as under-researched issues that we identified as relevant for LRC, such as domain bias, directionality, and semantic memorization. Moreover, from a linguistic point of view, adopting an underlying KG as source for our data, MUSCLE can help bridge the gap between paradigmatic and syntagmatic sense definitions (i.e., structuralist and post-structuralist approaches) (Storjohann, 2015) by allowing to extend the relations between lexemes with the surrounding context of each of them in the KG.

3. Dataset Design

In this section we detail the methodology followed to build the MUSCLE dataset, from the design requirements to the selection and extraction of data.

3.1. Dataset Requirements

To avoid the main limitations found in existing datasets, we set specific requirements for our own. In particular, MUSCLE must: (R1) cover as many areas of knowledge as possible; (R2) cover several lexico-semantic relation types; (R3) cover multiple natural languages; (R4) consist of subject–relation–object tuples where the relation can encode directional information (i.e., be asymmetric), so subject and object will not necessarily be interchangeable; (R5) define one natural language label for each subject, each relation, and each object appearing in the dataset for each of the languages considered; (R6) have a size at least comparable to those of previous studies; (R7) implement strategies to help avoid over-fitting and lexical memorization; (R8) finally, include Linked Open Data identifiers to establish unambiguous semantics and enable the enrichment of the dataset with external sources.

In the following, we detail how these requirements guided the development of the MUSCLE dataset.

3.2. Data Source and Extraction

We used Wikidata (wikidata.org), Wikimedia's peer-produced Knowledge Graph, as a data source.[4]

Structure. Most of Wikidata's content is organized in Items, each of which represents an entity in the world. Every Wikidata Item has a concept URI, optional labels in up to 566 possible languages (June 2023) to name the entity, and any number of statements about it. A statement connects two Items, the subject and the object (the object can also be a literal), via one Wikidata Property, which provides the semantics of the relation. Using Wikidata as a source allows us to easily meet requirements **R1**-**R4** and **R8**.

The entities represented by an Item are further classified into *classes* (categories or collections of individuals with common characteristics) and *instances* (individuals, concrete entities with characteristics that make them distinct from other concrete entities). We considered Items with subclass of (P279) statements as *classes* and those with instance of (P31) statements as *instances*. Note that there can be Items that are instances and classes at the same time. The MUSCLE dataset only includes classes, which usually have common names (e.g., human, city, painting, ...), leaving aside instances, which usually have proper names (e.g., Douglas Adams, New York City, Mona Lisa, ...).

Data Quality. Wikidata implements various quality systems combining different strategies, including a system of Property constraints to ensure high data quality, a machine learning system to label potentially harmful contributions called ORES[5], collaborative data models integrating Shape Expressions (ShEx) called Wikidata Schemas (Samuel, 2021), and manual review and correction processes typical of peer-production sites. Overall, contributions to Wikidata tend to align with information needs (Abián et al., 2022), achieving a quality level that has proven appropriate in a wide variety of scenarios, from Wikipedia infoboxes and Google KG applications to answers from virtual assistants such as Apple's Siri and Amazon's Alexa.

In addition, we followed some strategies to ensure the quality of the MUSCLE dataset:

1. We only considered *truthy statements*, those that had *"the best non-deprecated rank"* for each Property and subject Item.[6]

[4]We generated and used a Wikidata dump available at https://wdumps.toolforge.org/dump/3194.

[5]see https://www.wikidata.org/wiki/Wikidata:ORES

[6]Wikidata allows multiple data values (objects) for each subject-property pair, each value with one of three possible ranks: the default "normal" rank, the "preferred" rank (for the best and most current values), or the "deprecated" rank (for known errors or outdated information). By selecting only "truthy" statements, we opt for data values with the best non-deprecated rank for each subject-property pair, prioritizing "preferred" rank if available, or defaulting to "normal" rank otherwise. This approach helps filter out

2. We only considered classes whose content was expected to be more polished and agreed upon among more participants. We trivially achieved this by considering only the classes with labels defined in all the languages chosen for MUSCLE. Apart from assuring the quality of the data, this step fulfilled requirement **R3**.

3. We determined the type of lexico-semantic relation represented by each Property by aggregating expert judgments, fulfilling requirement **R2**.

4. We validated the quality of the MUSCLE dataset with the experiments described in Sections 4 and 6.

3.3. Property Selection

We included only Properties that: 1) had at least 4000 uses in Wikidata, 2) were used to state properties of a *class*, and 3) clearly represented one of the following types of lexico-semantic relations between a subject and an object: antonym of, meronym of, holonym of, hyponym of, hypernym of. Table 1 shows the nine selected Wikidata Properties. The selection was manually carried out by a team of five experts, consisting of linguists and knowledge engineers. They adhered to the definitions of the semantic relations outlined in (Storjohann, 2015). The selection process involved iterative discussions about a starting set of 700 Wikidata Properties[7] to establish clear mappings to Storjohann's categories. This process enabled the identification of a set of prototypical Properties while avoiding potentially controversial or spurious usages by prioritizing community adoption and consensus.

Despite being a limitation, we intentionally excluded synonymy from the MUSCLE dataset. Synonymy is considered one of the most complex and heterogeneous relations (Baroni and Lenci, 2011b), as it requires an understanding of the context of the target entities. Previous LRC datasets such as BLESS (Baroni and Lenci, 2011a) also avoid using synonyms due to their inherent complexity (they had difficulties in even finding convincing pairs for 200 concrete concepts). Note that, while there exist datasets that have kept this relation (such as EVALution), they have done so due to the nature of the sources they use: EVALution is created by filtering and combining ConceptNet 5.0 and WordNet 4.0, where the synonymy relation (accurate or not) is already established. Note, however, that such relation is not based at conceptual level, but rather at lexico-semantic level. Instead, as our approach is inspired by Semantic Maps, we build on the comparison and contraposition of verbalizations in different languages to define concepts by opposition (not having the synonymy relationship available by construction). Here, the nature of our dataset is crucial and shows that, from this point of view, it is complementary to previous existing datasets.

Moreover, the lexical relation does not align completely with the conceptual/logical notion of equivalence: logical equivalence is crisp, with both concepts needing to be completely the same, while synonymy can be argued to be gradual; thus, the alignment is far from trivial. In Wikidata, there are two different main properties to state the equivalence of two concepts: exactMatch (P2888) and said to be the same as (P460). The former one establishes links to other external datasets, which would be out of the scope of our requirements. The latter one is too noisy to extract synonyms, as each pair of related concepts would require manual validation. So, to ensure the quality of the dataset, the potential extraction of synonyms from Wikidata Items will be considered in future work.

Random Relation Generation. In order to include noisy relations between the terms and make the semantic classification more challenging, we exploit the nature of the graph to add relations between entities that are not related either lexically or semantically. We do this by adding random relations that are not actually present in the graph between concepts in the dataset, as well as mixing the orientation of the relation. We keep a proportion of random relations similar to previous datasets for the sake of comparison.

3.4. Language Selection

We selected the 40 most prevalent languages by word count in the dataset mix used to train GPT-3[8], from which we finally selected the 25 ones which had labels for all the concepts in the dataset, meeting requirements **R3** and **R5**. The selected languages are (in alphabetical order of their ISO 639-1 codes): Arabic (ar), Catalan (ca), Czech (cs), Danish (da), German (de), English (en), Spanish (es), Farsi (fa), Finnish (fi), French (fr), Hebrew (he), Hungarian (hu), Indonesian (id), Italian (it), Japanese (ja), Korean (ko), Dutch (nl), Polish (pl), Portuguese (pt), Russian (ru), Serbian (sr), Swedish (sv), Turkish (tr), Ukrainian (uk), and Chinese (zh).

4. Dataset Analysis

To detect different possible biases in our dataset, first, we analyzed the semantic domain of the included concepts (via community detection and taxonomy analysis) and, then, we sought for biases in the raw data using the Learning to Split approach (Bao and Barzilay, 2022).

Community Detection. Working with a knowledge graph as a starting point allows us to use community detection to analyze the semantic components of the dataset. This, in turn, allowed us to build a semantics-guided split-oriented to avoid semantic bias, as we will see in following sections. To analyze

lower-quality data. https://www.mediawiki.org/?oldid=5841902#Truthy_statements

[7]Those with more than 4000 uses, shared among 25 languages, not specific to Wikimedia and not deprecated.

[8]https://github.com/openai/gpt-3/blob/master/dataset_statistics/languages_by_word_count.csv

relation type	Wikidata Property	# uses in Wikidata	# uses in MUSCLE	semantic definition
antonym of	opposite of (P461)	27 320	296	symmetric
hyponym of (reversed, *hypernym of*)	subclass of (P279)	3 404 282	5 328	transitive, asymmetric
holonym of	has part(s) (P527)	2 144 459	1 360	transitive, inverse of part of (P361)
	made from material (P186)	1 494 634	243	
	has part(s) of the class (P2670)	41 989	76	asymmetric, subproperty of has part(s) (P527)
	contains (P4330)	9 761	4	
meronym of	part of (P361)	4 742 639	875	transitive, asymmetric, inverse of has part(s) (P527)
	part of the series (P179)	909 163	4	subproperty of part of (P361)
	anatomical location (P927)	4 420	4	subproperty of part of (P361)

Table 1: Wikidata Properties present in MUSCLE, along with their total uses in Wikidata (30 March 2023) and the types of lexico-semantic relations they represent.

the semantic communities in the dataset, we applied four different graph community detection algorithms[9]: weakly/strongly connected components (WCC/SCC), label propagation (LP), and the Louvain community detection method (Blondel et al., 2008). The graphs considered included the lexico-semantic relations (we excluded the *random* relation as it is just a random distractor). For all the algorithms, we ran two different setups: 1) a local configuration, taking into account only the concepts and relations included in the dataset; and 2) a global one, which uses all Wikidata classes to calculate the communities.

The first three algorithms (WCC, SCC, and LP) lead to very large components and did not provide insight into the structure of the KG. However, Louvain communities in the global configuration were smaller and exhibited a semantic cohesion that the others did not. Table 2 contains some of the detected communities and their domains (we include the complete communities in the repository dataset as well for further analysis). While the components were not semantically pure (i.e., some included concepts that could be considered as noise or being part of another subcommunity, e.g., *Politics*, *Linguistics*, or *Literature* are present in the biggest component), we could see that the smaller the component, the higher the semantic cohesion it exhibited. This enabled us to use these communities as a starting point to build a split aimed at avoiding semantic/domain memorization, as we will see in Section 5.

Hierarchical Information. Taking advantage of the KG nature of our data source, we also studied the semantic clustering achieved via the common ancestors of the concepts in the dataset. To do so, for each concept in our dataset, we gathered all its hierarchical parents that were not part of the dataset up to a depth of two (i.e., parent and grandparent concepts belonging to Wikidata but not to MUSCLE). Then, we calculated how many concepts in the dataset were descendants of such concepts to obtain the size of the semantic clusters.

[9]We used Neo4j's Graph DataScience (GDS) implementations of the algorithms.

domain	size	concept labels
Politics	168	political party, plutocracy, technocracy, kleptocracy, oligarchy, ...
Linguistics		lang. family, proto-language, constructed lang., dialect, idiolect, ...
Literature ...		creative work, written work, literary work, book series, letter, ...
Physics	107	physical quantity, potential energy, velocity, density, energy, ...
Occupations	86	writer, violinist, singer-songwriter, anthropologist, carpenter, ...
...	...	...
Music	38	indie rock, song, world music, alternative rock, popular music, ...
Medicine	27	dentistry, surgery, radiography, cardiology, endocrinology, ...
...	...	...
Disorders	10	autism, aphasia, tinnitus, intellectual disability, bulimia nervosa, ...
Noble titles	5	nobility, baron, princess, baronet, noble title
...	...	...

Table 2: Sample of Louvain-detected communities, along with the number of concepts in each.

Table 3 shows a sample of the broadest concepts. The differences in the concepts that were gathered assured the coverage of a broad range of domains. However, note that the semantic analysis of the domains was carried out only to discover possible memorization problems; there is room for further detailed analysis to characterize the semantics in the dataset in a finer-grained way. This is currently left as future work as the current analysis allowed us to perform the required splits in an informed way.

concept label	distance	# descendants
artificial entity	2	711
matter	2	252
human activity	2	107
...	...	...
mathematical concept	1	155
religion or world view	1	89
health problem	1	88
human language	1	61
music	1	54
...	...	...

Table 3: Sample of semantic groups detected via subclass of (P279), along with their distance (number of hops) and size in concepts within the MUSCLE dataset.

Bias Detection. For bias detection, we used the Learning to Split approach (Bao and Barzilay, 2022), which finds the dataset split that is most difficult to generalize. In such a split, prototypical elements are learned in the training split, while outlier elements are in the test split, making them more difficult to classify properly. Poor classification performance with such a split would point to a bias in the generalizability of the dataset. Some biases that we expected in the MUSCLE dataset were semantic bias (if all the elements of a semantic domain, e.g. sports, were predicted to have a particular semantic category) and morphological bias (if, for example, relations with words beginning with *in-*, which can be the prefix often used to produce antonyms, were always predicted as antonymic, even in cases like *inherit*. Yet, after inspecting the splits produced by the Learning to Split algorithm, we did not find any pattern that pointed to a particular bias.

5. Dataset Configurations

After the analysis presented in the previous section, we decided to provide two different train/test dataset splits, which will be used in the evaluation of the dataset itself:

- Random split (RanS): We performed a stratified random split (i.e., keeping the ratios of the different relations in the train/test sets) following a 50/50 train/test ratio[10].
- Semantic split (SemS): We split the dataset by semantic domains represented by the Louvain communities identified in our previous analyses. We added each community to a split and then discarded relations between splits through random exploration to semantically isolate the splits.

Table 4 shows the numbers of concepts and relations for each split. Note that, for SemS, there are no concepts shared between the train and test datasets. As MUSCLE covers 25 languages, the total numbers of different pairs are $343{,}825/343{,}850$ for RanS and $190{,}400/196{,}025$ for SemS, meeting requirement **R6**.

[10]Given the final size of the dataset, we consider that the training split can be further split to get a dev dataset.

Moreover, as we will see in Section 6.2, the SemS split meets requirement **R7**.

		RanS		SemS	
		train	test	train	test
concepts	split	6459	6464	3399	3414
	total	7231		6813	
relations	ant	148	148	145	123
	holo	841	842	470	691
	mero	473	472	340	354
	hyper	1332	1332	955	952
	hypo	1332	1332	993	888
	random	9627	9628	4713	4833
	total	13753	13754	7616	7841

Table 4: Numbers of concepts and relations in MUSCLE's train and test splits for RanS and SemS.

6. Dataset Evaluation

To gain further insight into the MUSCLE dataset, we designed and performed a set of experiments to answer the following research questions:

Q1 What performance is obtained by current methods when compared to other datasets for LRC? (Section 6.1)

Q2 Do the current dataset splits avoid the risks of lexical memorization? (Section 6.2)

Q3 Does including the directionality of the lexico-semantic relations in the dataset impact the performance of current methods? (Section 6.3)

Q4 Can the latest LLMs already perform these LRC tasks? (Section 6.4)

6.1. Comparison to Similar LRC Datasets

As a baseline, we adopted the minimal prompting approach proposed by Pitarch et al. (2023), the current SoTA for LRC using Pretrained Language Models (PLMs) and very simple prompts. As PLMs, we used RoBERTa (Liu et al., 2019) for English and XLM-R (Conneau et al., 2020) for multilingual experiments.

We ran two experimental setups: training with all the data in all the languages, and separating each language before training. In this experiment, we used three languages (namely, English, German, and Chinese) to compare the results with CogALex-VI (Karmakar and McCrae, 2020), the only previously existing multilingual dataset. We wanted to test how multilingual models benefit from witnessing the same relation in different languages in these datasets. Besides, as in CogALex-VI, since random pairs were added to the training set to increase the difficulty of the classification task, we excluded them when reporting the results.

Table 5 shows the results of both experiments (*-all refers to the results obtained for a particular language

dataset	lang	ant	holo	mero	hyper	hypo	macro avg not random	weighted avg not random
MUSCLE RanS	all	0.721	0.718	0.618	0.782	0.772	0.722	0.745
	en-all	0.738	0.729	0.623	0.789	0.773	0.731	0.751
	de-all	0.711	0.707	0.618	0.770	0.766	0.715	0.736
	zh-all	0.713	0.717	0.612	0.787	0.778	0.721	0.747
	en	0.734	0.720	0.631	0.789	0.784	0.731	0.753
	de	0.678	0.697	0.608	0.774	0.771	0.706	0.735
	zh	0.717	0.716	0.618	0.787	0.784	0.724	0.750
MUSCLE SemS	all	0.645	0.573	0.504	0.743	0.714	0.636	0.663
	en-all	0.677	0.610	0.519	0.746	0.728	0.656	0.680
	de-all	0.641	0.566	0.493	0.737	0.699	0.627	0.654
	zh-all	0.617	0.541	0.499	0.747	0.715	0.624	0.656
	en	0.404	0.515	0.277	0.713	0.698	0.521	0.599
	de	0.534	0.467	0.484	0.729	0.675	0.578	0.616
	zh	0.606	0.519	0.490	0.744	0.717	0.615	0.649
CogALex VI / 2.0	**lang**	**ant**	**hypo**		**syn**			
	all	0.727 / 0.788	0.605 / 0.676		0.584 / 0.667		0.639 / 0.711	0.641 / 0.713
	en-all	0.673 / 0.757	0.553 / 0.661		0.532 / 0.642		0.586 / 0.687	0.590 / 0.690
	de-all	0.679 / 0.735	0.548 / 0.607		0.508 / 0.594		0.578 / 0.645	0.579 / 0.646
	zh-all	0.940 / 0.956	0.847 / 0.863		0.852 / 0.874		0.880 / 0.898	0.882 / 0.900
	en	0.698 / 0.736	0.538 / 0.626		0.543 / 0.643		0.593 / 0.669	0.597 / 0.671
	de	- / -	- / -		- / -		- / -	- / -
	zh	0.909 / 0.909	0.805 / 0.805		0.792 / 0.792		0.836 / 0.836	0.839 / 0.839

Table 5: F1 score for MUSCLE, CogALex-VI and CogALex2.0. XLM-R fine-tuned in two setups: all languages against all languages, and each language on their own (English-en, German-de, and Chinese-zh). The results exclude the *random* relation. XLM-R trained only in German with CogALex-VI and CogALex2.0 did not converge.

having fine-tuned the model with all the languages)[11][12]. We can see that, using minimal prompting approach to fine-tune the model for LRC, the results are overall better on MUSCLE dataset than on CogALexVI 2.0 (focusing on the 'all' row where all languages are used for train and test, and the rightmost column, which reports the weighted average). However, we have to bear in mind that synonyms in CogaLexVI 2.0 are the most challenging category, while this category is not included in MUSCLE. Comparing the different languages separately, Pitarch et al. (2023) obtain similar results for both datasets except for Chinese, for which they obtain better results on CogALexVI.

We can see the effect of using all the languages compared to using just one in the training (*-all vs *), improving especially in SemS, CogALex-VI, and CogALex2.0 (where using only German does not even converge). These datasets pose a similar level of difficulty even though MUSCLE requires distinguishing the direction of the asymmetric relations. RanS shows more stable results regarding the training language. When comparing RanS and SemS results, we hypothesize that, apart from the training size, domain knowledge also plays a role in semantic classification, making SemS a more challenging dataset (similar to CogALex-VI and CogALex2.0, but without the pitfalls already analyzed and extending its multilinguality). Answering **Q1**, the performance is similar to the most challenging datasets and, as MUSCLE includes the direction of the relations, it makes LRC even more challenging (as shown in Section 6.3).

6.2. Checking for Memorization

Lexical memorization in LRC (Levy et al., 2015) occurs when some words systematically appear in a relation, predisposing a supervised model to learn that such particular words are signals of that relation, instead of learning the semantic relation between the source and target words. To measure the risk of memorization[13], we distinguish three types of tokens in a train/test dataset: 1) *indicators*, tokens in a source (target) word that mostly appear participating in the same relation in both train and test datasets; 2) *distractors*, tokens in a source (target) word that mostly appear participating in one relation in the train dataset, and in a different one in the test dataset; and 3) *independent elements*, tokens in the test dataset that are not in the train dataset. Their characterization can be parameterized by a threshold value β, establishing a minimum participation propor-

[11]The results for MUSCLE RansS and SemS trained with all languages and split by language can be consulted in AppendixB

[12]For the experiments conducted, we fine-tuned the large version of the RoBERTa and XLM-R language models using the Huggingface transformers library (Wolf et al., 2020) and the following setup: batch size of 32; AdamW optimizer with a learning rate, weight decay and warm-up ratio equal to 2e-5, 0.01, 0.01, respectively; and the models were trained during 6 epochs. Each experiment was run twice and the mean of the macro F1-score and the weighted F1-score by the support of the labels, without considering the random label, are reported. Training was performed on a Linux server with two A10 24GB GPUs. Overall, we consumed around 50 hours of GPU usage.

[13]We refer the interested reader to Appendix A.1 for the formal definition of all the metrics.

tion. For example, the token 'ham' is an indicator (distractor) among the source words for a threshold value $\beta = 0.7$, if there exists a relation label l such that 'ham' is participating in l 70% or more of its appearances among the source words of the train dataset, and the same (different) condition for 'ham' is met in the test dataset.

These elements allow us to define three risk metrics: 1) Risk of indicators (R_{ins}), the maximum proportion, between the source and target words, of word pairs in the test dataset containing only indicators; 2) Risk of distractors(R_{dis}), similar to the risk of indicators but containing only distractors; and 3) Risk of independent observations (R_{ind}), the proportion of word pairs in the test dataset composed of independent tokens. Our hypothesis is that the more *indicators* and the fewer *distractors* and *independent elements* a dataset contains, the greater the lexical memorization risk.

To validate it statistically, we calculated the risk values for 7 LRC datasets and gathered their results for at least 3 models (F1-scores) as presented in the literature. Table 6 shows these values using the multilingual-BERT pre-tokenizer and $\beta = 0.7$. For example, the first row in Table 6 reads as follows: for the K&H+N dataset, 75.0% of the observations in the test set contain only *indicators*, 0.8% contain only *distractors*, and 20.4% are composed of words that are not in the training set.

dataset	lang	$\mathbf{R_{ins}}$	$\mathbf{R_{dis}}$	$\mathbf{R_{ind}}$	$\overline{\mathbf{F1}}$
K&H+N		75.0	0.8	20.4	0.98
BLESS		65.2	1.5	20.1	0.93
ROOT09	en	31.2	13.3	16.5	0.88
EVALution		28.3	8.7	4.8	0.67
CogALexV		1.0	5.9	36.2	0.56
CogALexVI	en	2.6	25.3	23.2	0.60
	zh	13.7	12.8	12.1	0.90
	de	2.3	30.5	33.5	0.63
CogALex 2.0	en	3.3	21.1	30.3	0.79
	de	3.6	25.7	39.2	0.68
MUSCLE RanS	all	3.0	2.8	6.8	0.72
	en	3.1	2.7	6.5	0.73
	de	3.9	3.4	8.5	0.71
	zh	1.1	1.0	1.6	0.72
MUSCLE SemS	all	0.2	2.3	91.9	0.64
	en	0.2	2.6	93.3	0.64
	de	0.1	0.9	99.5	0.63
	zh	0.2	5.0	45.1	0.62

Table 6: Metrics for the risk of lexical memorization ($\beta = 0.7$) for the selected datasets and their reported F1 scores. MUSCLE results are obtained with the same setup as *-all configuration in Section 6.1.

The greater the lexical memorization risk of a dataset, the easier and better results would be expected with that dataset. To quantify this risk, we fitted a simple linear model of each risk measure against the collected results. Regarding the indicator risk R_{ins}, we observed a strong positive correlation with the results, with a positive Pearson correlation coefficient of 0.74. Regarding R_{dis} and R_{indp}, the correlation was slightly lighter but negative, with a negative Pearson correlation coefficient of -0.46 and -0.36, respectively, confirming the intuition that the more distractors and independent elements, the harder it is for a model to obtain better results. The p-values for all calculated regression coefficients were lower than 0.05, supporting our hypothesis statistically.

In Table 6, we also find the risk metrics ($\beta = 0.7$) for MUSCLE RanS and SemS for the full dataset and the English, German, and Chinese languages[14]. While MUSCLE RanS has a similar indicator risk (R_{ind}) to that of the CogALex datasets family (the most difficult ones so far), it has fewer distractors and independent values. Regarding MUSCLE SemS, we found a low indicator risk and a high level of independent words, making it in fact the most challenging dataset, as we will see in the next sections. Thus, we can affirmatively answer **Q2**.

6.3. Impact of Directionality

As other datasets for LRC do not include the direction of asymmetric relations (hyper-/hypo-, holo-/meronyms), to answer **Q3**, we analyzed the impact of not informing about the direction of the relations in MUSCLE by flattening such relations into non-directed ones.

Table 7 shows the results for the two versions of SemS (with/without considering the directions of the relations), using all languages for training and testing. We focused on SemS, since the previous experiments showed that it was more challenging and, given the hierarchical structure of the data source, domain transfer could interfere with the results, especially regarding holo- and meronymy relations (as shown in Table 5 - **all** rows).

Flattening the direction of the relations makes the task easier (higher values for mero-/holo- and hypo-/hypernymy relations). In turn, the classification of antonyms gets slightly affected, possibly because the model has less information about the structural differences (the direction seems to help discriminating among the five relations, even though it is more difficult to classify them). Thus, we can affirmatively answer **Q3**.

6.4. Can LLMs Already Perform the Task?

Given the recent developments with Large Language Models (LLMs), we evaluated the performance that ChatGPT (Brown et al., 2020) (on its version 3.5-turbo-0613) achieved with the MUSCLE dataset to answer question **Q4**. We used English as language and, as prompt, lists of subject-object pairs with blanks in the middle preceded by the following zero-shot instruction (up to ten pairs were submitted to ChatGPT per query):

Fill in the blank of each item with one of the following options: "hyponym of", "hyperonym of", "meronym of", "holonym of", "antonym of", or "unrelated to":

"dog" __ "animal".

Table 8 shows the results achieved with this prompt and those obtained with minimal prompting using RoBERTa, for comparison's sake. While ChatGPT performs quite well for antonyms, it fails with directional relations, at least in this zero-shot training scenario.

[14]We include in Appendix A.2 the metrics for all languages using different tokenizers.

dataset	lang	ant	holo	mero	hyper	hypo	macro avg not random	weighted avg not random
MUSCLE SemS	all	0.587	0.496	0.449	0.682	0.666	0.576	0.603
		ant	**mero/holo**		**hypo/hyper**		–	–
		0.545	0.504		0.727		0.592	0.642

Table 7: F1 score for MUSCLE SemS. Model trained and tested using all languages. The results exclude the *random* relation. Results when flattening the asymmetric relations in the lower row.

	model	training	ant	holo	mero	hyper	hypo	macro avg not rand	weighted avg not random
MUSCLE SemS	RoBERTa	fine-tunning	0.715	0.615	0.689	0.776	0.761	0.578	0.709
	ChatGPT	zero-shot	0.626	0.160	0.175	0.383	0.466	0.362	0.342

Table 8: F1 score for monolingual experiments in English to compare to LLMs.

The F1 scores for the "unrelated to" category were 0.918 for RoBERTa and 0.912 for ChatGPT, which rules out this category as the source of ChatGPT's poor results. Thus, answering **Q4**, we can foresee that some prompt-tuning and further research must be done to adapt LLMs to the out-of-context LRC task.

7. Conclusions and Future Work

In this work, after identifying some limitations in existing datasets for computational Lexical Relation Classification (LRC), we have proposed a novel dataset aiming to solve these shortcomings and foster the research in the field by bridging lexical and conceptual semantics in Language Models.

The MUSCLE dataset exploits a well-known KG, providing a solid semantic grounding and a crowdsourced-curated multilingual representation in 25 languages. Moreover, we have analyzed the data from different angles to avoid lexical and semantic biases, providing two different splits that present several difficulties. Last but not least, the dataset distinguishes the asymmetry of the lexico-semantic relations.

As future work, we will consider extending the dataset to other NLP tasks by exploiting the linked nature of Wikidata. For example, we will follow and search for further lexico-semantic information, including OntoLex-LEMON data (McCrae et al., 2017).

Acknowledgements

Supported by the Spanish project PID2020-113903RB-I00 (AEI/FEDER, UE), by DGA/FEDER, by the Agencia Estatal de Investigación of the Spanish Ministry of Economy and Competitiveness and the European Social Fund through the "Ramón y Cajal" program (RYC2019-028112-I), and by DGA Government predoctoral fellowship.

Limitations

Prototypicallity: Providing concepts with lexical forms available in 25 languages constrains the inclusion of concepts with lexical forms in only a few languages, which can lead to lexical and cultural gaps. Our dataset includes prototypical concepts and excludes long-tail concepts.

Completeness of relations: We have mapped nine Wikidata Properties as semantic relations to their lexical counterparts. The chosen semantic relations are the most prominent ones, but we are aware that more semantic relations could be mapped to the five lexical relations, so we will assess and possibly include them in future versions.

Synonyms: As mentioned in the body of the paper, we have discarded the synonym relation due to the misalignment between the notions of synonymy and the conceptual equivalence, and the noise we witnessed in the most similar Wikidata Property, said to be the same as (P460). We acknowledge that is a limitation of the current dataset, but the difficulty of defining such a relation properly hinders the possible quality of the data extracted from Wikidata. We will study the possibility of extending MUSCLE using Linguistic Linked Open Data to cover synonymy.

Quality of translations: We must acknowledge possible limitations in the language translations, as we have not been able to completely assess their quality. Those translations in the languages that the authors read and speak (English, French, Spanish, German) seem to be correct, but the authors do not master all 25 languages covered. Finding people for all remaining languages is a very difficult task that was out of our possibilities. Anyway, it should be considered that the quality of our dataset is proportional to the quality of Wikidata: If Wikidata is well annotated, so our dataset is

Multilingual usage: The MUSCLE dataset is intended to be used in a multilingual setting, as it is where Semantic Map theory can be applied. However, its multilingual orientation also provides benefits in comparison to previous datasets for LRC, including the analysis and control of memorization, semantic domain distribution, generalizability, and relation directionality.

8. Bibliographical References

David Abián, Albert Meroño-Peñuela, and Elena Simperl. 2022. An analysis of content gaps versus user needs in the Wikidata Knowledge Graph. In *The Semantic Web – ISWC 2022*, pages 354–374, Cham. Springer International Publishing.

Alfred V. Aho and Jeffrey D. Ullman. 1972. *The Theory of Parsing, Translation and Compiling*, volume 1. Prentice-Hall, Englewood Cliffs, NJ.

American Psychological Association. 1983. *Publications Manual*. American Psychological Association, Washington, DC.

Rie Kubota Ando and Tong Zhang. 2005. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6:1817–1853.

Galen Andrew and Jianfeng Gao. 2007. Scalable training of L_1-regularized log-linear models. In *Proceedings of the 24th International Conference on Machine Learning*, pages 33–40.

Yujia Bao and Regina Barzilay. 2022. Learning to split for automatic bias detection. *arXiv preprint arXiv:2204.13749*.

Oren Barkan, Avi Caciularu, and Ido Dagan. 2020. Within-between lexical relation classification. In *EMNLP 2020 - 2020 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*.

Marco Baroni, Raffaella Bernardi, Ngoc Quynh Do, and Chung Chieh Shan. 2012. Entailment above the word level in distributional semantics. In *EACL 2012 - 13th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings*.

Marco Baroni and Alessandro Lenci. 2011a. How we blessed distributional semantic evaluation. In *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*.

Marco Baroni and Alessandro Lenci. 2011b. How we blessed distributional semantic evaluation. In *GEometrical Models of Natural Language Semantics*.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623. ACM.

Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008.

Gemma Boleda. 2019. Distributional semantics and linguistic theory. *ArXiv*, abs/1905.01896.

Zied Bouraoui, Jose Camacho-Collados, and Steven Schockaert. 2020. Inducing relational knowledge from bert. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7456–7463.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Jose Camacho-Collados, Luis Espinosa-Anke, and Steven Schockaert. 2020. Relational word embeddings. In *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*.

Ashok K. Chandra, Dexter C. Kozen, and Larry J. Stockmeyer. 1981. Alternation. *Journal of the Association for Computing Machinery*, 28(1):114–133.

Emmanuele Chersoni, Giulia Rambelli, and Enrico Santus. 2016. Cogalex-v shared task: Root18. In *Proceedings of the 5th Workshop on Cognitive Aspects of the Lexicon (CogALex - V)*, pages 98–103.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

James W. Cooley and John W. Tukey. 1965. An algorithm for the machine calculation of complex Fourier series. *Mathematics of Computation*, 19(90):297–301.

Alexandre François. 2008. Semantic maps and the typology of colexification: Intertwining polysemous networks across languages. *From polysemy to semantic change: Towards a typology of lexical semantic associations*, 106:163.

Thanasis Georgakopoulos and Stéphane Polis. 2018. The semantic map model: State of the art and future avenues for linguistic research. *Language and Linguistics Compass*, 12(2):1–33.

Dan Gusfield. 1997. *Algorithms on Strings, Trees and Sequences*. Cambridge University Press, Cambridge, UK.

Martin Haspelmath. 2003. The geometry of grammatical meaning: Semantic maps and cross-linguistic comparison. *New Psychology of Language, v.2, 211-242 (2003)*, 2.

Marti A. Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In *International Conference on Computational Linguistics*.

Aidan Hogan, Eva Blomqvist, Michael Cochez, Claudia d'Amato, Gerard de Melo, Claudio Gutiérrez, Sabrina Kirrane, José Emilio Labra Gayo, Roberto Navigli, Sebastian Neumaier, Axel-Cyrille Ngonga Ngomo, Axel Polleres, Sabbir M. Rashid, Anisa Rula, Lukas Schmelzeisen, Juan F. Sequeda, Steffen Staab, and Antoine Zimmermann. 2021. Knowledge graphs. *ACM Computing Surveys*, 54(4):71:1–71:37.

Saurav Karmakar and John P. McCrae. 2020. Cogalex-vi shared task: Bidirectional transformer based identification of semantic relations. In *Proceedings of the Workshop on the Cognitive Aspects of the Lexicon*, pages 65–71. ACL.

Christian Lang, Lennart Wachowiak, Barbara Heinisch, and Dagmar Gromann. 2021. Cogalex 2.0: Impact of data quality on lexical-semantic relation prediction. In *NeurIPS Data-Centric AI Workshop*.

Omer Levy, Steffen Remus, Chris Biemann, and Ido Dagan. 2015. Do supervised distributional methods really learn lexical inference relations? In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 970–976, Denver, Colorado. Association for Computational Linguistics.

Hugo Liu and Push Singh. 2004. Conceptnet — a practical commonsense reasoning tool-kit. *BT Technology Journal*, 22:211–226.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*.

John P McCrae, Julia Bosque-Gil, Jorge Gracia, Paul Buitelaar, and Philipp Cimiano. 2017. The ontolex-lemon model: development and applications. In *Proceedings of eLex 2017 conference*, pages 19–21.

Ken McRae, George S. Cree, Mark S. Seidenberg, and Chris McNorgan. 2005. Semantic feature production norms for a large set of living and nonliving things. *Behavior Research Methods*, 37:547–559.

George A. Miller. 1995. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41.

M.L. Murphy. 2003. *Semantic Relations and the Lexicon: Antonymy, Synonymy, and Other Paradigms*. Cambridge University Press.

Silvia Necsulescu, Sara Mendes, David Jurgens, Núria Bel, and Roberto Navigli. 2015. Reading between the lines: Overcoming data sparsity for accurate classification of lexical relationships. In *Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics*, pages 182–192. Association for Computational Linguistics.

Lucia Pitarch Ballesteros, Jorge Bernad, Lacramioara Dranca, Carlos Bobed, and Jorge Gracia. 2023. No clues, good clues: Out of context lexical relation classification. In *ACL 2023 - 61th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, page To appear.

Mohammad Sadegh Rasooli and Joel R. Tetreault. 2015. Yara parser: A fast and accurate dependency parser. *Computing Research Repository*, arXiv:1503.06733. Version 2.

Christoph Rzymski, Tiago Tresoldi, SimonGreen, Hill, Mei-Shin Wu, Nathanael E. Schweikhard, Mária, Koptjevskaja-Tamm, Volker Gast, Timotheus A. Bodt, Abbie Hantgan, Gereon A. Kaiping, Sophie Chang, Yunfan, Lai, Natalia Morozova, Heini Arjava, Nataliia Hübler, Ezequiel Koile, Steve Pepper, Mariann Proos, Briana, Van Epps, Ingrid Blanco, Carolin Hundt, SergeiMon, akhov, Kristina Pianykh, Sallona Ramesh, Derek Russell, Gray, Robert Forkel, and Johann-Mattis List. 2020. The database of cross-linguistic colexifications, reproducible analysis of cross-linguistic polysemies. *Scientific Data*, 7.

John Samuel. 2021. Shexstatements: Simplifying shape expressions for wikidata. In *Companion Proceedings of the Web Conference 2021*, WWW '21, page 610–615, New York, NY, USA. Association for Computing Machinery.

Enrico Santus, Anna Gladkova, Stefan Evert, and Alessandro Lenci. 2016a. The cogalex-v shared task on the corpus-based identification of semantic relations. In *Proceedings of the Workshop on Cognitive Aspects of the Lexicon*, pages 69–79.

Enrico Santus, Alessandro Lenci, Tin-Shing Chiu, Qin Lu, and Chu-Ren Huang. 2016b. Nine features in a random forest to learn taxonomical semantic relations. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4557–4564, Portorož, Slovenia. European Language Resources Association (ELRA).

Enrico Santus, Frances Yung, Alessandro Lenci, and Chu-Ren Huang. 2015. Evalution 1.0: an evolving semantic dataset for training and evaluation of distributional semantic models. In *Proceedings of the 4th Workshop on Linked Data in Linguistics: Resources and Applications*, pages 64–69. Association for Computational Linguistics.

Carina Silberer and Mirella Lapata. 2012. Grounded models of semantic representation. In *Proceedings of the 2012 Joint Conference on Empirical Methods*

in Natural Language Processing and Computational Natural Language Learning, pages 1423–1433, Jeju Island, Korea. Association for Computational Linguistics.

Suzanne Stevenson and Paola Merlo. 2022. Beyond the benchmarks: Toward human-like lexical representations. *Frontiers in Artificial Intelligence*, 5.

Petra Storjohann. 2015. Sense relations. In Nick Riemer, editor, *The Routledge Handbook of Semantics (1st Ed.)*. Routledge.

Simone Tedeschi, Johan Bos, Thierry Declerck, Jan Hajic, Daniel Hershcovich, Eduard H. Hovy, Alexander Koller, Simon Krek, Steven Schockaert, Rico Sennrich, Ekaterina Shutova, and Roberto Navigli. 2023. What's the meaning of superhuman performance in today's nlu? In *ACL 2023 - 61th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, page To appear.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Commun. ACM*, 57:78–85.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: A free collaborative knowledgebase. *Commun. ACM*, 57(10):78–85.

Chengyu Wang, Minghui Qiu, Jun Huang, and Xiaofeng He. 2021. Keml: A knowledge-enriched meta-learning framework for lexical relation classification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35:13924–13932.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Rong Xiang, Emmanuele Chersoni, Luca Iacoponi, and Enrico Santus. 2020. The CogALex shared task on monolingual and multilingual identification of semantic relations. In *Proceedings of the Workshop on the Cognitive Aspects of the Lexicon*, pages 46–53, Online. Association for Computational Linguistics.

A. Lexical Memorization Risk Metrics

Lexical memorization (Levy et al., 2015) in LRC (Levy et al., 2015) appears when there are words that occur systematically in a relation, enabling a supervised model to learn that some particular words are signals of a particular relation, instead of learning the characteristics of the semantic relation itself. Thus, for example, for words that represent broad categories such as `animal`, a model can learn that when `animal` appears as target word, then a hyponymy relation occurs no matter what the source word is. In fact, not only words can be memorized, but also punctuation marks or any other type of strings. For instance, if we have samples such as (`object-oriented programming, programming, hyponym`) or (`white-collar worker,employee,hyponym`), the model could learn that '–' is an indicator of the hyponym relation. Moreover, since language models such as BERT, RoBERTa or the family of the GPT models, work based on a pre-trained tokenizer, these models can memorize that the apparition of a specific token is an indicator of a relation.

Thus, we can talk about a more general problem than lexical memorization, this is, *token memorization*, and we propose metrics to measure the exposure of a dataset to the risk of detecting it. Note that when tokens are directly words, then we would be dealing with lexical memorization.

One simple way to detect token memorization is would be to test the model with a dataset that has no tokens in common between train and test splits. However, this strategy is very restrictive: on the one hand, since the tokenizers of the current language models split texts into a moderate small set of tokens, it could be difficult to find train/test splits without tokens in common; and, on the other hand, this strategy prevents checking whether the model can learn different relations between two same words from those seen during training.

A.1. Formal Definition

Given a dataset split into train/test, we want to measure if the test split is good enough to detect the token memorization that can occurs while training a model using such train split.

Formally, we consider that a dataset $\mathcal{D}=\{D^r,D^e\}$ is composed of two sets, D^r and D^e, named *train and test sets*, respectively. A train or test set D^i is a finite set of triples (observations), $D^i=\{(s,t,l)\,|\,s,t\in\Sigma^*,\,l\in L\}$ where s and t are strings over an alphabet Σ and $L=\{1,...,K\}$ is a set of integer labels. The strings s and t are called *source* and *target* strings, respectively. We denote by D^i_{Src} and D^i_{Tgt} the set of all source and target strings in D^i. For sake of simplicity in notation, we also denote by $D^i=D^i_{Src}\cup D^i_{Tgt}$. A tokenizer tok is a function that splits any string w into substrings called tokens, $tok(w)=(s_1,...,s_m)$. By an abuse of the notation, we also consider that $tok(w)$ is the set of the tokens composing w. For a set M of strings, $tok(M)$ denotes the set of all tokens in strings of M, $tok(M)=\cup_{w\in M}tok(w)$. Given a train or test set D^i and a string $w\in tok(D^i_{Src})$, we denote by $\mathbf{p}_w^{Src_i}\in[0,1]^K$ to the distribution vector containing the observed proportions in $tok(D^i_{Src})$ of the string w participating in the j-th relation ($j\in L$, the set of possible relation labels), corresponding to the j-th positions of the vector. Similarly, we denote by $\mathbf{p}_w^{Tgt_i}$

the distribution when we consider that $w \in tok(D^i_{Tgt})$. $\mathbf{p}_w^{D^i_{Src}}$, $\mathbf{p}_w^{D^i_{Tgt}}$ are *the source and target label distributions* of the string w in D^i. Then, we can define:

Definition 1 *A string w is said to be a β-*source indicator*, $\beta \in [0,1]$, for the dataset $\mathcal{D} = \{D^r, D^e\}$, if it holds the following conditions:*

$$\begin{cases} \max(\mathbf{p}_w^{D^r_{Src}}) > \beta \wedge \max(\mathbf{p}_w^{D^e_{Src}}) > \beta \\ \text{argmax}(\mathbf{p}_w^{D^r_{Src}}) = \text{argmax}(\mathbf{p}_w^{D^e_{Src}}) \end{cases} \quad (1)$$

Definition 2 *A string w is said to be a β-*source distractor*, if it holds the following conditions:*

$$\begin{cases} \max(\mathbf{p}_w^{D^r_{Src}}) > \beta \wedge \max(\mathbf{p}_w^{D^e_{Src}}) > \beta \\ \text{argmax}(\mathbf{p}_w^{D^r_{Src}}) \neq \text{argmax}(\mathbf{p}_w^{D^e_{Src}}) \end{cases} \quad (2)$$

We define a β-target indicator and distractor in the same way as the previous definitions but using the Tgt sets. Finally, we have:

Definition 3 *A string w is* independent *if:*

$$w \in tok(D^e) \setminus tok(D^r)$$

We can extend these definitions to the observations (s,t,l) included in the dataset as follows:

Definition 4 *Given a dataset $\mathcal{D} = \{D^r, D^e\}$, an observation in the test dataset $(s,t,l) \in D^e$ is said to be a β-source indicator observation if for all $w \in tok(s)$, w is a β-source indicator.*

Analogously, we define a β-target indicator, β-source distractor, β-target distractor, and an independent observation. We will denote by β-p^{Src}_{ins}, β-p^{Tgt}_{ins}, β-p^{Src}_{dis}, β-p^{Tgt}_{dis}, p^D_{ind}, the percentage of observations in the test set that are β-source indicators, β-target indicators, β-source distractors, β-target distractors, and independent observations, respectively.

Finally, we define three metrics to measure the risk of token memorization of a dataset.

Definition 5 *Given a dataset $\mathcal{D} = \{D^r, D^e\}$, a tokenizer tok, $\beta \in [0,1]$ and the previous definitions, we define the* risk of indicators, distractors and independent observations*, respectively, as:*

- $R_{ins} = \max(\beta\text{-}p^{Src}_{ins}, \beta\text{-}p^{Tgt}_{ins})$
- $R_{dis} = \max(\beta\text{-}p^{Src}_{dis}, \beta\text{-}p^{Tgt}_{dis})$
- $R_{ind} = p_{ind}$

A.2. Risk Metrics for MUSCLE

In this subsection, we include all the risk values obtained for MUSCLE RanS and SemS splits considering all the different languages covered. We present the results for three different tokenizers: multilingual-BERT pre-tokenizer (Table 9), XML-R (Table 10), and GPT-3.5 tokenizers (Table 11).

	MUSCLE RanS			MUSCLE SemS		
Lang.	$\mathbf{R_{ins}}$	$\mathbf{R_{dis}}$	$\mathbf{R_{ind}}$	$\mathbf{R_{ins}}$	$\mathbf{R_{dis}}$	$\mathbf{R_{ind}}$
all	3.0	2.8	6.8	0.2	2.3	91.9
ar	2.9	2.6	6.5	0.3	3.1	92.9
ca	3.2	2.6	6.9	0.3	2.4	93.4
cs	3.6	3.2	7.9	0.0	1.9	98.0
da	3.5	3.3	8.3	0.2	1.1	99.3
de	3.9	3.4	8.5	0.1	0.9	99.5
en	3.1	2.7	6.5	0.2	2.6	93.3
es	3.0	2.9	6.9	0.3	2.1	94.0
fa	3.1	2.5	6.7	0.1	2.2	94.0
fi	3.7	3.3	8.5	0.2	0.9	99.2
fr	3.1	2.7	6.9	0.4	2.5	93.9
he	3.3	2.9	7.0	0.1	2.8	95.1
hu	3.4	3.1	8.1	0.1	1.3	98.7
id	2.5	2.4	6.3	0.2	2.7	91.8
it	3.0	2.7	7.1	0.3	1.7	94.2
ja	1.7	1.7	4.7	0.3	3.6	71.3
ko	3.5	3.1	8.0	0.2	1.7	97.8
nl	3.5	3.2	8.4	0.1	1.1	98.7
pl	3.2	3.0	7.9	0.1	1.6	97.7
pt	3.1	2.8	7.0	0.2	2.2	93.7
ru	3.4	3.1	7.9	0.1	2.3	97.9
sr	3.6	3.1	7.8	0.1	2.3	98.1
sv	3.6	3.2	8.5	0.3	1.6	99.4
tr	3.4	2.7	7.1	0.2	2.7	96.0
uk	3.5	3.1	7.7	0.2	1.8	98.1
zh	1.1	1.0	1.6	0.2	5.0	45.1

Table 9: Risk metrics for MUSCLE RanS and SemS using multilingual-BERT pre-tokenizer and $\beta = 0.7$.

	MUSCLE RanS			MUSCLE SemS		
Lang.	$\mathbf{R_{ins}}$	$\mathbf{R_{dis}}$	$\mathbf{R_{ind}}$	$\mathbf{R_{ins}}$	$\mathbf{R_{dis}}$	$\mathbf{R_{ind}}$
all	1.7	1.5	2.0	0.2	7.1	53.9
ar	2.5	1.1	2.1	0.2	6.6	52.9
ca	2.5	1.9	4.2	0.3	5.3	78.6
cs	3.2	2.1	4.7	0.3	5.7	84.4
da	2.7	2.2	5.1	0.2	5.7	87.2
de	3.3	1.9	4.3	0.2	6.7	83.2
en	2.5	2.0	4.5	0.2	4.5	84.3
es	2.9	2.1	4.6	0.3	5.3	82.1
fa	1.6	1.1	1.7	0.2	10.0	50.1
fi	2.6	2.0	4.4	0.2	6.1	81.3
fr	2.2	1.9	4.1	0.4	5.5	78.9
he	1.7	0.9	1.3	0.4	6.8	36.5
hu	2.2	2.2	4.0	0.2	8.4	77.1
id	2.4	2.4	4.5	0.3	5.5	84.2
it	2.2	2.1	4.6	0.2	5.0	81.5
ja	2.4	1.0	1.7	0.2	8.4	46.3
ko	0.8	0.7	1.3	0.2	5.2	32.5
nl	2.7	2.0	4.8	0.1	5.1	84.3
pl	2.5	1.9	4.5	0.1	5.8	80.6
pt	2.8	1.9	4.6	0.3	5.6	82.3
ru	3.2	2.1	3.7	0.3	6.9	77.7
sr	2.3	1.7	3.5	0.2	6.5	75.9
sv	2.4	2.2	4.8	0.2	6.8	83.5
tr	3.0	2.1	4.2	0.2	6.5	80.4
uk	2.7	1.8	3.1	0.3	8.7	72.1
zh	2.0	1.6	2.9	0.2	5.3	69.4

Table 10: Risk metrics for MUSCLE RanS and SemS using XLM-R tokenizer and $\beta = 0.7$.

Lang.	MUSCLE RanS			MUSCLE SemS		
	$\mathbf{R_{ins}}$	$\mathbf{R_{dis}}$	$\mathbf{R_{ind}}$	$\mathbf{R_{ins}}$	$\mathbf{R_{dis}}$	$\mathbf{R_{ind}}$
all	0.7	0.7	1.0	0.1	3.1	26.0
ar	0.0	0.0	0.0	0.0	0.1	0.4
ca	2.4	1.6	3.5	0.4	5.1	71.2
cs	1.3	1.6	3.3	0.1	5.3	64.4
da	2.6	1.9	4.4	0.2	5.1	76.7
de	2.7	1.7	3.6	0.1	7.7	74.7
en	3.0	2.6	5.6	0.2	3.1	91.5
es	2.6	1.5	3.5	0.3	5.0	72.7
fa	0.0	0.0	0.0	0.0	0.0	0.7
fi	2.1	1.5	3.0	0.1	5.2	62.4
fr	2.3	1.9	4.1	0.2	6.3	74.5
he	0.0	0.1	0.1	0.0	0.0	2.2
hu	1.2	1.4	2.8	0.1	6.2	55.6
id	2.0	1.6	3.1	0.2	5.9	66.5
it	2.3	1.4	3.5	0.4	5.9	67.9
ja	0.8	0.2	0.3	0.1	1.6	8.7
ko	0.0	0.1	0.0	0.0	0.5	3.4
nl	1.4	1.9	4.1	0.2	6.7	76.8
pl	1.4	1.3	3.0	0.1	7.1	60.9
pt	2.3	1.5	3.6	0.3	6.3	70.5
ru	0.1	0.1	0.4	0.0	0.8	8.2
sr	0.1	0.3	0.8	0.0	1.2	12.2
sv	2.1	1.9	3.9	0.1	5.8	71.2
tr	1.9	1.2	3.0	0.2	7.5	59.7
uk	0.1	0.2	0.3	0.0	1.9	7.7
zh	0.1	0.1	0.3	0.1	1.9	9.2

Table 11: Risk metrics for MUSCLE RanS and SemS using GPT-3.5 tokenizer and $\beta = 0.7$.

B. Complete Experiments

We include the results of all the values obtained for the experiments presented in Section 6.1 with all the languages for MUSCLE splits. Tables 12 and 13 contain the values for the adopted minimal prompting approach (Pitarch Ballesteros et al., 2023), training with all the languages and aggregating the results for each of them.

	lang	ant	holo	mero	hyper	hypo	macro avg not random	weighted avg not random
MUSCLE RanS	all	0.700	0.701	0.574	0.750	0.738	0.693	0.714
	ar	0.687	0.658	0.528	0.720	0.696	0.658	0.676
	ca	0.700	0.706	0.572	0.737	0.730	0.689	0.709
	cs	0.697	0.707	0.581	0.756	0.734	0.695	0.717
	da	0.707	0.707	0.580	0.764	0.751	0.702	0.725
	de	0.705	0.712	0.590	0.758	0.759	0.705	0.728
	en	0.727	0.726	0.593	0.781	0.759	0.717	0.739
	es	0.700	0.699	0.583	0.760	0.738	0.696	0.718
	fa	0.703	0.681	0.547	0.727	0.708	0.673	0.690
	fi	0.683	0.700	0.585	0.749	0.744	0.692	0.716
	fr	0.722	0.709	0.576	0.757	0.744	0.701	0.721
	he	0.647	0.678	0.544	0.719	0.703	0.658	0.683
	hu	0.735	0.699	0.586	0.757	0.749	0.705	0.722
	id	0.660	0.703	0.580	0.752	0.753	0.690	0.719
	it	0.726	0.709	0.580	0.756	0.742	0.702	0.721
	ja	0.697	0.707	0.573	0.741	0.729	0.689	0.709
	ko	0.649	0.681	0.541	0.723	0.723	0.663	0.691
	nl	0.716	0.704	0.593	0.766	0.749	0.706	0.726
	pl	0.716	0.704	0.588	0.752	0.741	0.700	0.719
	pt	0.705	0.706	0.584	0.760	0.746	0.700	0.722
	ru	0.732	0.706	0.588	0.759	0.746	0.706	0.724
	sr	0.712	0.700	0.571	0.741	0.724	0.690	0.707
	sv	0.683	0.718	0.578	0.761	0.750	0.698	0.725
	tr	0.660	0.690	0.545	0.740	0.729	0.673	0.701
	uk	0.728	0.701	0.568	0.757	0.750	0.701	0.721
	zh	0.697	0.712	0.589	0.764	0.758	0.704	0.729

Table 12: F1 score for MUSCLE RanS using the minimal prompting approach.

	lang	ant	holo	mero	hyper	hypo	macro avg not random	weighted avg not random
MUSCLE SemS	all	0.587	0.496	0.449	0.682	0.666	0.576	0.603
	ar	0.500	0.465	0.413	0.623	0.597	0.519	0.549
	ca	0.486	0.481	0.440	0.646	0.650	0.541	0.579
	cs	0.570	0.504	0.450	0.696	0.684	0.581	0.614
	da	0.601	0.457	0.449	0.693	0.682	0.576	0.603
	de	0.600	0.523	0.449	0.710	0.682	0.593	0.624
	en	0.627	0.531	0.474	0.731	0.710	0.615	0.644
	es	0.644	0.485	0.471	0.689	0.682	0.594	0.612
	fa	0.544	0.518	0.459	0.648	0.639	0.562	0.589
	fi	0.616	0.458	0.421	0.670	0.665	0.566	0.588
	fr	0.596	0.499	0.464	0.684	0.664	0.581	0.606
	he	0.533	0.448	0.431	0.652	0.616	0.536	0.564
	hu	0.557	0.498	0.458	0.704	0.677	0.579	0.614
	id	0.583	0.512	0.472	0.691	0.692	0.590	0.620
	it	0.605	0.488	0.449	0.661	0.665	0.573	0.595
	ja	0.622	0.537	0.455	0.689	0.652	0.591	0.613
	ko	0.523	0.528	0.425	0.657	0.628	0.552	0.586
	nl	0.606	0.506	0.434	0.698	0.660	0.581	0.608
	pl	0.609	0.505	0.461	0.681	0.681	0.588	0.612
	pt	0.608	0.471	0.465	0.666	0.660	0.574	0.594
	ru	0.612	0.544	0.473	0.707	0.695	0.606	0.635
	sr	0.631	0.437	0.428	0.679	0.649	0.565	0.583
	sv	0.581	0.457	0.452	0.688	0.679	0.571	0.600
	tr	0.565	0.449	0.389	0.648	0.642	0.539	0.567
	uk	0.626	0.513	0.455	0.698	0.672	0.593	0.616
	zh	0.620	0.566	0.494	0.728	0.720	0.626	0.656

Table 13: F1 score for MUSCLE SemS using the minimal prompting approach.