Towards Sense to Sense Linking across DBnary Languages

Gilles Sérasset
Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG
GETALP team
38000 Grenoble, France
gilles.serasset@imag.fr

Abstract

Since 2012, the DBnary project extracts lexical information from different Wiktionary language editions (26 editions in 2025) and makes it available to the community as queryable RDF data (modelled using the Ontolex-lemon ontology). This dataset contains more than 12M translations linking languages at the level of Lexical Entries. This paper presents an effort to automatically link the DBnary languages at the Lexical Sense level. For this, we explore different ways to compute cross-lingual semantic similarity, using multilingual language models.

1 Introduction

Even in the era of Large Language Models pretrained in unsupervised settings, Lexical Resources (LR) are still in use and have proved useful for advancing various natural language processing (NLP) tasks. For instance, such resources may enhance the quality of machine translation by providing accurate cross-lingual mappings, thus improving translation fidelity (Jones et al., 2023). They are also of importance for end users that easily refer to them through on-line browsing or mobile dictionary apps.

Since 2012, the DBnary dataset[1] extracts lexical data from 26 Wiktionary language editions[2] and makes it available as an RDF dataset. Extracted from one of the most important community built lexical resource, it contains lexical entries in many languages, along with definitions, lexico-semantic relations, translation, among other lexical information. One of the shortcomings of the current dataset lies in the lack of semantic alignment between language editions.

The final objective of this work will be twofold: **1.** providing cross-lingual links at the lexical sense level, based on the translations available at the entry or surface form level, and **2.** associating each DBnary lexical sense with an embedding in a unique multilingual vector space.

In this paper, we explore the use of Multilingual Neural Language Models for the computation of a cross-lingual semantic similarity measure that we use to align existing translation pairs at the semantic level. After describing the current way DBnary dataset models cross-lingual links (section 2), we will define the task at hand and the related work we borrow from (section 3), then describe a gold standard dataset we built to evaluate different approaches (section 3.2). We proceed with the experiments (section 4) and results (section 5) and discuss shortcomings of the approaches for the systematic modelling of translations at the semantic level and for the distribution of sense embeddings to the end users (section 6).

2 The DBnary Dataset

DBnary (Sérasset, 2012, 2015) is a large multilingual lexical dataset extracted from 26 language editions of the Wiktionary project. It is made available following the Lexical Linked Open Data principles using Ontolex-lemon model in RDF format, following Chiarcos et al. (2011). Overall, this dataset describes 7.9M Lexical Entries, accounting for 6.5M Lexical Senses, usually described with a textual definition. Additionally, it contains 12.3M translation pairs.

The **OntoLex**[3] model, is a community standard for machine-readable lexical resources that has been adopted by many data providers for its ability to ensure FAIR principles,[4] and the dominant

[1] https://kaiko.getalp.org/about-dbnary

[2] Just like Wikipedia, there are different editions of Wiktionary (that differ by their URL (e.g http://en.wiktionary.org refers to the English edition and http://fr.wiktionary.org to the French edition). Following Meyer and Gurevych (2012), we call each of these independent web sites a *language edition*.

[3] https://www.w3.org/2016/05/ontolex

[4] FAIR (for *Findable Accessible Interoperable and Reusable*) refers to a set of principles dedicated to allow for re-use of any research object.

vocabulary for modeling machine-readable dictionaries as Linked Data. The goal of OntoLex is to represent lexical resources as a knowledge graph, allowing the integration of information from different dictionaries and to facilitate the exchange, storage, and reusability of lexical information. The Ontolex-lemon model is a W3C community report consisting of a core model, along with additional modules (mainly, **lime** for metadata, **synsem** for the description of syntax and semantics, **decomp** for decomposition of terms into subterms and **vartrans** to represent lexico-semantic and translation relations).

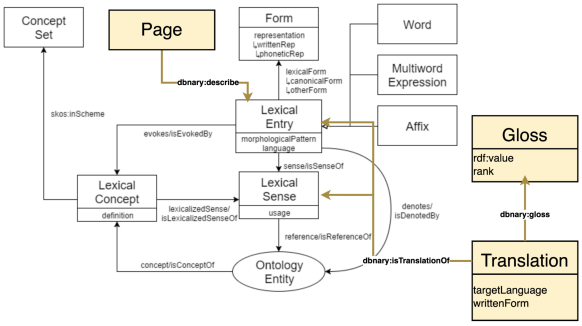
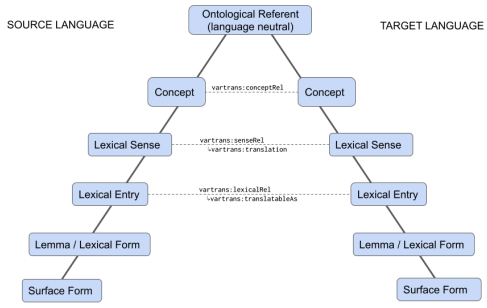


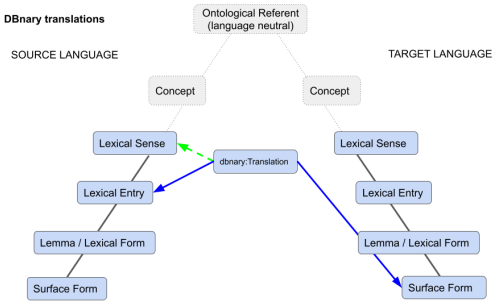
Figure 1: The Ontolex model along with DBnary extension used to represent Wiktionary pages and translations.

The atomic unit of information of Wiktionary is a *page*, where several lexical *entries* may be described. Such entries usually share their canonical form (which usually corresponds to the name of the page). The organisation of entries in the page and the structure and content of such entries differ according to the Wiktionary language editions, but usually contain definitions of the senses and a set of lexical information (etymology, morphology, lexico-semantic derivations, ...), along with translations in other languages. As an example, the cat_{eng}[5] page in the English edition describes 6 entries (3 nouns, 2 verbs and 1 adjective). Other lexical entries in other languages are also described and extracted in DBnary; however, for this article, we will focus only on the lexical entries of the language edition (what we call the *endolex*).

DBnary uses Ontolex core model to represent the extracted data, hence, it gives a common representation of lexical data, despite the differences in the way such lexical information is organised in each edition. Ad-hoc classes were added to be able to represent pages (`dbnary:Page`) and also to represent translations pairs (`dbnary:Translation`), along with glosses (`dbnary:Gloss`) that are often used to provide users with context to disambiguate the lexical sense for which a lexical information stands. The DBnary resulting model is shown in Figure 1.



(a) The **vartrans** module allows for the representation of cross-lingual links either at the Lexical Entry level or at the Lexical Sense level.



(b) The Wiktionary data is often insufficient to decide if the target of a translation is a Lexical Entry in the target language, and, if it is, which one (in case of homonymy), also, there are usually no information allowing to decide on the target sense that is involved in a translation.

Figure 2: A graphical representation of the cross-lingual linking strategies, according to the semantic level of the links and inspired by the Vauquois triangle (Vauquois, 1968) according to Gracia et al. (2025).

The Ontolex model defines the **vartrans** extension to encode relations in general and translations in particular. However, DBnary authors chose a nonstandard representation. The reason is illustrated in Figure 2. Figure 2a shows that the **vartrans** module may be used either to link two Lexical Entries or two Lexical Senses together. However, in Wiktionary, each language edition is independent of the other, and the available translations are given as strings, with no guarantee that they correspond to a lexical entry in the target edition. For example, $persignar_{cat}$ is translated in English as "cross oneself on forehead, lips, and heart", which does not correspond to a valid English lexical entry.

[5]In this article we will use $entry_{lg}$ to denote the page named "*entry*" in the edition for language *lg*, this particular page is available at `https://en.wiktionary.org/wiki/cat`

Also, since Wiktionary is an ongoing collaborative dictionary, the target page may not exist while being a perfectly valid lexical entry in the target lexicon. Finally, even if the page exists, it is not possible to systematically decide which lexical entry is the target of the translation. For all these reasons, DBnary decided to reify each translation pair using `dbnary:Translation` class where the source lexical entry is usually known and the target is represented as a surface form rather than as a link to a lexical entry in the target lexicon (in blue in Figure 2b).

Many translations are also associated with their source lexical sense (in dotted green in Figure 2b) that are selected using Tchechmedjiev et al. (2014) with an accuracy of 0.82 to 0.96 F1 score.

3 Linking Language Editions at the Lexical Sense Level

The task we address in this paper is the following: "How can we efficiently identify the correct source and target lexical sense(s) that are involved in available translation pairs?"

The main objective of this work is to include such links using the **vartrans** module in DBnary and, as a by-product, provide the multilingual lexical sense embeddings that allowed for this linking. With 12M translations and more than 6M lexical senses with definitions that continuously change while the language communities collaboratively correct and expand the editions, DBnary extracts a new version twice a month. So, links must be recomputed at each extraction, and the efficiency of the method should be assessed both in terms of performance and in terms of frugality in computing resources.

3.1 Related Work

This work follows on Tchechmedjiev et al. (2014), which attempted to identify the lexical sense of translation sources by leveraging monolingual similarity measures using a two-level string distance based on Tversky index (Tversky, 1977) (sentence similarity distance computed on a sequence of tokens, with token similarity computed with a character-level string distance).

In this initial work, approaches using statistical measures like Jimenez et al. (2012) were disregarded as they were requiring too much computation times for the statistical model computation in a multilingual setting where the number of languages was growing. However, times have changed, and today, many language models are available for semantic similarity measure computation.

The task at hand implies being able to compute the similarity between lexical sense definitions in different languages. For this, we tried several strategies.

3.1.1 Token similarity measure

The first strategy modified the two-level similarity measure used in Tchechmedjiev et al. (2014) by substituting the token-level character-based string distance with cosine similarity between *fastText* non contextual token embeddings (Bojanowski et al., 2016) trained and aligned on multilingual texts (Joulin et al., 2018).

3.1.2 Sentence similarity measures

Many models are now trained to directly compute sentence similarity. For the monolingual task, we could use monolingual models; however, in multilingual settings we need the model to be multilingual and compute similarity between sentences in different languages. In this work, we focus on multilingual sentence similarity models, as the final objective is to align definitions in as many language pairs as possible. We evaluated Multilingual Universal Sentence Encoder (MUSE) (Yang et al., 2020), Language-agnostic BERT sentence embedding (LaBSe) (Feng et al., 2022), Language-Agnostic SEntence Representations (LASER) (Artetxe and Schwenk, 2019a,b), Sentence-Level Multimodal and Language-Agnostic Representations (SONAR) (Duquenne et al., 2023), Multilingual E5 Text Embeddings (Wang et al., 2024), mGTE (Zhang et al., 2024), and original multilingual pretrained models from sentence BERT (Reimers and Gurevych, 2019, 2020).[6] OpenAI text embedding models (via openai API) (Neelakantan et al., 2022) have also been used for comparison purposes only.

3.1.3 Machine Translation based similarity measures

Another approach for computing the cross-lingual similarity measure is to rely on a machine translation system to compute similarity on two texts in the same language. In this work, we used Opus-MT

[6] Namely `paraphrase-multilingual-mpnet-base-v2` (paraphrase), `stsb-xlm-r-multilingual` (stsb), `static-retrieval-mrl-en-v1` (static) and `static-similarity-mrl-multilingual-v1` (static-similarity).

models from Tiedemann et al. (2023) and Tiedemann and Thottingal (2020) to translate target languages into English before computing monolingual similarity with sentence similarity measures.

3.2 Sense to Sense Linking Gold Standard

We evaluated the different strategies on a gold standard dataset created from translation pairs and lexical sense definitions extracted from DBnary. To create this dataset we extracted 96 English pages from DBnary, chosen among frequent and highly ambiguous English terms. These pages described 232 different lexical entries and 2646 different English lexical senses.

Translation pairs from English to Chinese, French, German, Italian, Russian and Spanish were extracted. For this experiment, we only selected translations that were associated to a textual gloss helping to disambiguate the source sense. Table 1a shows the resulting number of entries and translation pairs per Part of Speech.

For each translation pair, we extracted target lexical entries and lexical senses. Each target lexical entry was also associated to a fake lexical sense *[NAWS]* (Not A Word Sense) created to identify lexical entries that are a valid translation target, but for which none of the described lexical senses was a valid target. Table 2 shows an example of the resulting data that is presented to the annotators.

Six annotators identified the source and target sense(s) involved in each translation pair, given the associated gloss. This implied two successive tasks: (1) *Monolingual task*: selecting the English definition associated to the English gloss, and (2) *Cross-lingual task*: selecting the target definition(s) associated to the translation pair, given the English selected sense definition. The annotator may select more than one target sense if necessary, and if no target sense is to be found, select *[NAWS]* definition for the appropriate lexical entry.

Agreement is measured using Kripendorff alpha (Krippendorff, 2025) on both tasks. For monolingual task $\alpha = 0.966$, which is considered high agreement, while $\alpha = 0.674$ for cross-lingual task, which is just above the α value (0.667) considered as minimal for data to be used to draw tentative conclusions. This agreement value is coherent with previous observation drawn when creating word sense disambiguation (WSD) datasets with fine-grain word sense definitions (Véronis, 1998; Murray and Green, 2004) as it is the case here.

After cleanup, systematic errors correction, and majority vote for disagreeing annotations, the dataset contains 2927 annotated translation pairs. Table 1b gives the number of annotated translation per target language. The annotated dataset is available at https://gricad-gitlab.univ-grenoble-alpes.fr/getalp/dbnary-translations-disambiguation

	N.	**Vb.**	**Adj.**	**Adv.**	**Int.**
entries	116	82	30	3	1
pairs	1711	907	263	25	3

(a) Repartition of part of speech for English source entries and translation pairs.

deu	**fra**	**ita**	**rus**	**spa**	**zho**
873	622	639	147	530	98

(b) Number of annotated translation pairs per target language.

Table 1: Insights on the annotated dataset built for the task.

4 Experiments

4.1 Monolingual Task

The purpose of the monolingual task is to identify the source sense that is denoted by a gloss associated to a translation pair. The first column of table 2 shows the set of English definitions among which to chose the one that is referred to by the gloss "*device made of flexible material*". In Wiktionary, glosses are usually a shorter version of the intended lexical sense definition. In our example, the gloss refers to lexical sense *5.4* of spring_{eng} noun.

In a first approach, one could try to select the most similar definition in the set of available definitions D, based on a similarity measure (Sim) with the gloss g (Equation 1), hence selecting only one lexical sense per gloss.

$$\arg\max_{d \in D} \text{Sim}(d, g) \tag{1}$$

However, in practice, the gloss may refer to several lexical senses of the entry. Hence, we allow the selection of multiple senses if their similarity is within a window size δ (Equation 2).

$$\left\{ d \in D \mid \text{Sim}(d, g) \geq \max_{d' \in D} \text{Sim}(d', g) - \delta \right\} \tag{2}$$

We chose to address three strategies for the computation of the similarity measure.

spring$_{eng}$**, noun**	*muelle*$_{spa}$
1. (countable) An act of springing: a leap, a jump.	**adj. 1**
2. (countable) The season of the year in temperate regions in which plants spring from the ground and into bloom and dormant animals spring to life.	**1.** Delicado, suave, blando.
2.1. (astronomy) The period from the moment of vernal equinox (around March 21 in the Northern Hemisphere) to the moment of the summer solstice . . . "	**2.** Voluptuoso.
2.2. (meteorology) The three months of March, April, and May in the Northern Hemisphere and September, October, and November in the Southern Hemisphere.	NAWS
3. (uncountable, figurative) The time of something's growth; the early stages of some process.	**sust. 1**
3.1. (figurative, politics) a period of political liberalization and democratization	**1.** Pieza elástica, usualmente de metal, colocada de modo que pueda utilizarse la fuerza que hace para recobrar su posición natural cuando ha sido separada de ella.
4. (countable, fashion) Someone with ivory or peach skin tone and eyes and hair that are not extremely dark, seen as best suited to certain colors of clothing.	**2.** Adorno compuesto de varios relicarios o dijes, que las mujeres de distinción llevaban pendiente a un lado de la cintura.
5. (countable) Something which springs, springs forth, springs up, or springs back, particularly	**3.** En plural| Tenazas grandes que se usan en las casas de moneda para agarrar los rieles y tejos durante la fundición y echarlos en la copela.
5.1. (geology) A spray or body of water springing from the ground.	NAWS
5.2. (oceanography, obsolete) The rising of the sea at high tide.	**sust. 2**
5.3. (oceanography) Short for spring tide, the especially high tide shortly after full and new moons.	**1.** Náutica.| Obra de piedra, hierro o madera, construida en dirección conveniente en la orilla del mar o de un río navegable, y que sirve para facilitar el embarque y . . .
5.4. A mechanical device made of flexible or coiled material that exerts force and attempts to spring back when bent, compressed, or stretched.	**2.** Transporte.| Andén alto, cubierto o descubierto, que en las estaciones de ferrocarriles sirve para la carga y descarga de mercancías.
5.5. (nautical) A line from a vessel's end or side to its anchor cable used to diminish or control its movement.	NAWS
5.6. (nautical) A line laid out from a vessel's end to the opposite end of an adjacent vessel or mooring to diminish or control its movement.	
5.7. (figurative) A race, a lineage.	
5.8. (figurative) A youth.	
5.9. A shoot, a young tree.	
5.10. A grove of trees; a forest.	
6. (countable, slang) An erection of the penis.	
7. (countable, nautical, obsolete) A crack which has sprung up in a mast, spar, or (rare) a plank or seam.	
8. (uncountable) Springiness: an attribute or quality of springing, springing up, or springing back, particularly	
8.1. Elasticity: the property of a body springing back to its original form after compression, stretching, etc.	
8.2. Elastic energy, power, or force.	
9. (countable) The source from which an action or supply of something springs.	
10. (countable) Something which causes others or another to spring forth or spring into action, particularly	
10.1. A cause, a motive, etc.	
10.2. (obsolete) A lively piece of music.	
NAWS	

Table 2: Example of the *spring*$_{eng}$ to *muelle*$_{spa}$ translation pair which is associated with the gloss: "*device made of flexible material*". presented with this extract, the annotator has to select the correct word sense in English (monolingual task) and in Spanish (cross-lingual task) taking the gloss into account.

We first reproduced the results from Tchechmedjiev et al. (2014) as a baseline to compare with the other similarity measures.

Then, we evaluated token similarity measure (see 3.1.1) borrowing the sentence similarity computation from Tchechmedjiev et al. (2014) and replacing the token similarity computation with cosine similarity on aligned fastText vectors.

Our third strategy uses sentence similarity measures (see 3.1.2), with cosine similarity on sentence embedding models that compute an unique vector for each definition.

For better interpretation of the results, we also provide two heuristics that were frequently used in WSD tasks: 1. random selection of a word sense and 2. systematic selection of the word sense described first in the lexical entry.

4.2 Multilingual Task

The input of the multilingual task is a source sense (supposedly identified by the monolingual task) and a surface form in a target language. The surface form is used to query DBnary for Lexical Entries on their canonical form (lemma) and their associated Lexical Senses. A pseudo sense labelled "*NAWS*" is added to each lexical entry.

The purpose of the task is to choose the lexical senses that are the adequate target for this specific translation of the given source lexical sense. Also, if none of the given target lexical senses are adequate, the multilingual task should identify the target lexical entry by selecting its associated "*NAWS*" pseudo sense.

For example, the second column of table 2 shows the lexical entries and senses associated to the Spanish translation of *spring*$_{eng}$ sense *5.4*. Among those, the task should choose sense *1.* of the first nominal entry of *muelle*$_{spa}$.

The approach used in this task is borrowed from the monolingual task Equation 2, where the gloss is substituted by the source sense definition and the similarity measures are multilingual.

The approach should also be able to decide that none of the lexical senses are fit to be selected

as targets and should choose the fallback "*NAMS*" pseudo word sense. For this, we introduce the Ω hyper-parameter which is the minimum similarity for which a lexical sense is eligible to be selected (Equation 3).

$$\text{let } m = \max_{d' \in D} \text{Sim}(d', g) \text{ in } \left\{ d \in D \mid \begin{array}{l} \text{Sim}(d, g) \geq m - \delta \quad \& \\ \text{Sim}(d, g) \geq \Omega \end{array} \right\} \tag{3}$$

Tchechmedjiev et al. (2014) cannot be used in this setting as the string distance used as token similarity measure is not multilingual. Token and sentence similarity measures (see 3.1.1 and 3.1.2) may be used provided that the models embed token or sentences into the same vector space regardless of the token or sentence language. We also experimented with Machine Translation based similarity measures (see 3.1.3).

4.3 Hyper-parameters Optimisation

All hyper-parameters were optimised using a grid search on 20% of the evaluation data. All results below are computed using optimised hyper-parameters.

Among the hyper-parameters, δ represents the ability of the method to select more than one target word-senses if their respective similarity are close enough, while Ω represents the ability of the model to decide that the most similar target sense is a valid choice or if none of the target senses are valid.

Of course, these hyper-parameters depend on the model, however, their values and behaviours follow common tendency. As an example Figure 3 shows the hyper-parameter influence for the sentence embeddings *paraphrase* model.

The source code for the experiments is available at `https://gricad-gitlab.univ-grenoble-alpes.fr/getalp/sense2sense_translations`.

5 Results

All tasks are evaluated using standard set-matching metrics, i.e. Recall, Precision and F1 score. All scores are computed using the optimal hyper-parameters setting.

5.1 Monolingual Task

Table 3 shows the performance of the different approaches. For better understanding of the task, we evaluated the heuristic consisting in selecting the first sense as the predicted answer that is frequently used in WSD tasks, based on the hypothesis that Wiktionary senses are given in order of usage frequency. Unlike usual WSD tasks where the first sense is also the most frequent sense (hence the most frequent answer), this heuristic is not significantly better than the random baseline.

Model	F1	precision	recall
random	0.131	0.131	0.131
first-sense	0.214	0.214	0.214
tchechmedjiev	0.914	0.908	0.929
fasttext	0.899	0.890	0.921
e5-instruct	0.899	0.899	0.899
gte-multilingual-base	0.928	0.928	0.928
labse	0.901	0.882	0.946
laser	0.763	0.763	0.763
muse	0.889	0.875	0.923
paraphrase	0.930	0.930	0.930
sonar	0.783	0.760	0.843
static	0.929	0.922	0.944
static-similarity	0.932	0.926	0.946
stsb	0.877	0.877	0.877
text-embedding-3-large	**0.946**	**0.946**	0.946
text-embedding-3-small	<u>0.944</u>	0.936	**0.962**
text-embedding-ada-002	<u>0.944</u>	0.944	0.944

Table 3: Results of the monolingual task, with F1, precision and recall scores. Bold values are maximum, and non significantly different scores are underlined.

In Tchechmedjiev et al. (2014) the reported F1-scores were 0.826 for French, 0.865 for Portuguese and 0.968 for Finnish. The evaluation was not available for English, due to the way the gold standard was (automatically) generated using glosses that were both given as a short text summarising a definition and as a word sense number (that were taken as the ground truth). The original string distance based monolingual word sense identification performs significantly better when evaluated on our gold standard than when evaluated using the original automatic gold standard generation. The reason for this difference comes from the fact that Wiktionary is an ever changing resources, and when word senses are edited, added, removed or re-ordered, the numerical glosses that refer to them are sometimes not updated and become out of sync with the set of definitions. Hence the original performance of this method was underestimated and one can see that it outperforms many sentence similarity models despite its very efficient computational cost.

The best-performing models are the OpenAI embedding models accessible through the OpenAI

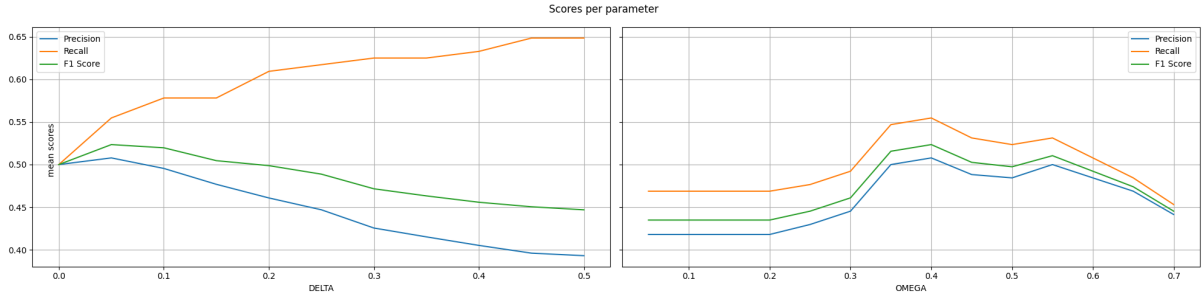


Figure 3: Influence of the hyper-parameters on the system performance for the *paraphrase* embedding in English to Italian task. In most models, $\delta = .05$, which shows it is a good choice to slightly take the risk to accept several lexical sense as valid targets, while Ω is much more fluctuant and highly depend on the model.

API. However, it is unclear what the actual energy cost of such models is, and using these models for the entire DBnary is likely to incur a cost that is not sustainable in our context. Moreover, some open-source model results may be considered almost as good as those of the OpenAI models, even if they are significantly lower.

5.2 Multilingual Task

Table 4 presents the results of the multilingual task, using F1 score. The random and first-sense baselines are provided.

For this task, OpenAI models also give the best results. However, we chose not to rely on these models in our use cases for reasons explained in section 6. For this reason, OpenAI models are only evaluated for reference, and are not included in best result and significance computations. These models are given in an independent table section were maximum values and significances are computed independently.

Results show that no model wins it all. The performance of sentence embeddings depends on the target language. Using a translation step plus monolingual sentence embedding seems to give slightly better results, but has a higher energy cost as the translation cost is added to embedding cost. However, this approach could allow for the use of a better monolingual sentence embedding model. Further evaluation is needed on this aspect.

6 Discussions and Limitations

Monolingual evaluation shows that it is still difficult to surpass Tchechmedjiev et al. (2014) which is based solely on string distance computation, with a very small energy cost for all languages. In this task, the added value of some models may not justify the energy cost. Thanks to the manually built gold standard, we also showed that the original results were underestimated.

In multilingual settings, sentence embedding models yield the best scores. However, it is not entirely clear whether the overall performance is sufficient to create links of high quality for inclusion in a distributed dataset. More detailed analysis should be conducted to understand what makes this task so challenging. The structure of definitions (frequently structured as genus-differencia) may differ significantly from the structure of sentences used to train the sentence embedding models. Moreover, lexical senses of the same entry are expected to share most of their semantics, and the task at hand should focus more on their differences rather than on their similarities.

OpenAI models delivered the best results in both monolingual and multilingual tasks. However, we cannot rely on these embeddings for our use case. The first reason is a general concern that arises when OpenAI models or chat services are used in a research setting. The fact that OpenAI embeddings, architectures, and datasets are closed source and largely unexplained prevents us from gaining any understanding of the reasons behind their success, nor can we determine whether the success on the gold standard is generalisable to real-world data. Although we are relatively confident that our (newly created) dataset was not part of those model training, we also have a particular reason to exclude OpenAI from our work.

Our objective is twofold: 1) to provide sense-to-sense translation links and 2) to distribute the embeddings associated with each lexical sense, so that end users can compute similarities at the lexical sense level. For the first objective, the best model should be used if its energy cost is reasonable on the scale of the full DBnary dataset. How-

	deu	fra	ita	rus	spa	zho	Average
random	0.307	0.222	0.254	0.329	0.238	0.510	0.310
first-sense	0.391	0.383	0.328	0.522	0.358	0.544	0.421
fasttext	0.444	0.511	0.511	0.558	0.502	0.513	0.507
translation+fasttext	0.446	0.412	0.368	0.427	0.400	0.526	0.430
e5-instruct	0.513	0.595	0.393	0.652	0.472	0.609	0.539
gte-multilingual-base	0.535	0.585	0.569	0.595	0.521	**0.614**	0.570
labse	0.507	0.524	0.503	0.632	0.539	0.595	0.550
laser	0.441	0.435	0.483	0.468	0.443	0.560	0.472
muse	0.511	0.545	0.527	0.641	0.551	0.609	0.564
paraphrase	0.570	**0.611**	0.514	0.663	0.597	0.591	0.591
sonar	0.516	0.519	0.510	0.649	0.494	0.597	0.547
static	0.329	0.355	0.447	0.236	0.327	0.339	0.339
static-similarity	0.485	0.529	0.549	0.540	0.535	0.605	0.540
stsb	0.520	0.593	0.545	**0.675**	0.546	0.588	0.578
translation+e5-instruct	0.504	0.577	0.380	0.584	0.487	0.593	0.521
translation+gte-multilingual-base	0.527	0.595	0.555	0.610	0.547	0.604	0.573
translation+labse	0.497	0.509	0.501	0.574	0.522	0.611	0.536
translation+laser	0.425	0.421	0.496	0.469	0.404	0.543	0.460
translation+muse	0.529	0.557	0.556	0.640	0.577	**0.614**	0.579
translation+paraphrase	**0.570**	0.604	0.523	0.628	**0.617**	0.611	**0.592**
translation+sonar	0.498	0.530	0.513	0.591	0.471	0.586	0.531
translation+static	0.536	0.549	0.537	0.563	0.539	0.569	0.549
translation+static-similarity	0.498	0.571	0.568	0.604	0.561	0.565	0.561
translation+stsb	0.529	0.576	**0.574**	0.603	0.569	0.599	0.575
text-embedding-3-large	*0.606*	*0.667*	0.588	*0.693*	0.623	*0.677*	*0.642*
text-embedding-3-small	0.591	0.616	0.594	0.629	*0.629*	0.646	0.617
text-embedding-ada-002	0.539	0.621	0.405	0.601	0.494	0.590	0.542
translation+text-embedding-3-large	0.588	0.639	*0.611*	0.625	0.624	0.619	0.618
translation+text-embedding-3-small	0.581	0.644	0.590	0.657	0.612	0.605	0.615
translation+text-embedding-ada-002	0.538	0.616	0.407	0.598	0.496	0.636	0.548

Table 4: F1 measure (higher is better) for optimal hyper-parameters for each language and averaged over languages. Maximum scores are given in bold and values that do not differ significantly (i.e. when p-values > .05) from best results are underlined. In all but latest section, OpenAI models are disregarded for maximum and significance computation. The latest section gather results using OpenAI API text embedding models. In this section, maximums are given in italics and significance is computed taking OpenAI models into account.

ever, providing embeddings tied to a closed-source, proprietary model would tie DBnary users to the model provider for their own use case. This would force any user of the DBnary embeddings to pay OpenAI to exploit them in their use cases.

Finally we should also note two limitations of this preliminary work. The first one comes from the dataset that only shows translation from English to other languages. Other sources should be added to the dataset. Moreover, more entries should be added to get more Chinese and Russian translations in order to have more significant results.

The second limitation comes from the main methodology. In this preliminary work, we take the glosses as a starting point for sense-to-sense link computation. The monolingual task identify the source lexical sense from the gloss, then the multilingual task is performed using the identified source lexical sense to identified the target one. This means that we can only deal with translations that are associated to a gloss. In English this represents 96.8% of the available translations. For all other translations, the sense-to-sense cross-lingual link cannot be computed with this process and 3.2% of the available translations will be disregarded (see Appendix A for full statistics on the availability of glosses in all DBnary extracted languages). Among all DBnary languages, the proportion of translations that are associated to a textual gloss are very imbalanced from 0% to 99.6%. As an example, with this methodology we will only handle 34.6% of the translations of French words.

7 Conclusion

This preliminary study is a first step towards a better cross-lingual link modelling in the DBnary dataset. Although some results are encouraging, additional work is required to achieve our goal. Focussing on our first objective (identifying lexical senses involved in translation pairs), definition embeddings bring a lot but will certainly benefit of other approaches.

If we succeed in computing a cross-lingual semantic similarity measure between DBnary lexical senses, we will be able to provide such cross-lingual links, but also distribute embeddings for each lexical sense in the dataset. Such embeddings could be used to query for semantically related senses and could be reused by end-users in downstream tasks. Many applications would benefit from these: browsing DBnary with direct access to close senses, bootstrapping/aligning models with lexical sense embedding rather than embeddings associated to surface forms, or linking such senses with several ontologies.

However, more work should be done before achieving results that are good enough for the computed data to be distributed along with the original DBnary data. For this, we need to further study the similarity measures we may apply for such definitions that could take into account the specificities of definition and better discriminate between definitions that share much semantics but differ on specific aspects.

Other solutions should also be investigated to handle all definitions that are not associated with a source gloss, in order to benefit from the richness of the original Wiktionary data.

Acknowledgements This research has been done in the context of the Cost Action CA23147 - Global Network on Large-Scale, Cross-domain and Multilingual Open Knowledge Graphs (GOBLIN).

8 Bibliographical References

References

Mikel Artetxe and Holger Schwenk. 2019a. Margin-based parallel corpus mining with multilingual sentence embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3197–3203, Florence, Italy. Association for Computational Linguistics.

Mikel Artetxe and Holger Schwenk. 2019b. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*.

Christian Chiarcos, Sebastian Hellmann, and Sebastian Nordhoff. 2011. Towards a linguistic linked open data cloud: The open linguistics working group. *Traitement Automatique des Langues*, 52(3):245–275.

Paul-Ambroise Duquenne, Holger Schwenk, and Benoît Sagot. 2023. Sonar: Sentence-level multimodal and language-agnostic representations. *Preprint*, arXiv:2308.11466.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Jorge Gracia, Gilles Sérasset, Michael Rosner, Ilan Kernerman, and Katerina Gkirtzou. 2025. Cross-lingual linking representation levels on the web of data. Submitted to NLP journal, preprint at `https://zenodo.org/`.

Sergio Jimenez, Claudia Becerra, and Alexander Gelbukh. 2012. Soft Cardinality: A Parameterized Similarity Function for Text Comparison. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*.

Alex Jones, Isaac Caswell, Ishank Saxena, and Orhan Firat. 2023. Bilex rx: Lexical data augmentation for massively multilingual machine translation. *Preprint*, arXiv:2303.15265.

Armand Joulin, Piotr Bojanowski, Tomas Mikolov, Hervé Jégou, and Edouard Grave. 2018. Loss in translation: Learning bilingual word mapping with a retrieval criterion. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

Klaus Krippendorff. 2025. *Content Analysis: An Introduction to Its Methodology*, fourth edition edition. SAGE Publications, Inc., Thousand Oaks, California. Especially chapter 12.

Christian M. Meyer and Iryna Gurevych. 2012. Wiktionary: A new rival for expert-built lexicons? Exploring the possibilities of collaborative lexicography. In Sylviane Granger and Magali Paquot, editors, *Electronic Lexicography*, chapter 13, pages 259–291. Oxford: Oxford University Press.

G.Craig Murray and Rebecca Green. 2004. Lexical knowledge and human disagreement on a wsd task. *Computer Speech & Language*, 18(3):209–222. Word Sense Disambiguation.

Arvind Neelakantan, Tao Xu, Raul Puri, Alec Radford, Jesse Michael Han, Jerry Tworek, Qiming Yuan, Nikolas Tezak, Jong Wook Kim, Chris Hallacy, Johannes Heidecke, Pranav Shyam, Boris Power, Tyna Eloundou Nekoul, Girish Sastry, Gretchen Krueger, David Schnurr, Felipe Petroski Such, Kenny Hsu, and 6 others. 2022. Text and code embeddings by contrastive pre-training. *Preprint*, arXiv:2201.10005.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525, Online. Association for Computational Linguistics.

Gilles Sérasset. 2012. Dbnary: Wiktionary as a LMF based Multilingual RDF network. In *Language Resources and Evaluation Conference, LREC 2012*, Istanbul, Turkey. Nicoletta Calzolari and Khalid Choukri and Thierry Declerck and Mehmet Uğur Doğan and Bente Maegaard and Joseph Mariani and Jan Odijk and Stelios Piperidis.

Gilles Sérasset. 2015. DBnary: Wiktionary as a Lemon-Based Multilingual Lexical Resource in RDF. *Semantic Web – Interoperability, Usability, Applicability*, 6(4):355–361.

Andon Tchechmedjiev, Gilles Sérasset, Jérôme Goulian, and Didier Schwab. 2014. Attaching Translations to Proper Lexical Senses in DBnary. In *3rd Workshop on Linked Data in Linguistics: Multilingual Knowledge Resources and Natural Language Processing*, pages 5–12, Reykjavik, Iceland.

Jörg Tiedemann, Mikko Aulamo, Daria Bakshandaeva, Michele Boggia, Stig-Arne Grönroos, Tommi Nieminen, Alessandro Raganato Yves Scherrer, Raul Vazquez, and Sami Virpioja. 2023. Democratizing neural machine translation with OPUS-MT. *Language Resources and Evaluation*, 58(2):713–755.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT — Building open translation services for the World. In *Proceedings of the 22nd Annual Conferenec of the European Association for Machine Translation (EAMT)*, Lisbon, Portugal.

Amos Tversky. 1977. Features of Similarity. *Psychological Review*, 84(2):327–352.

Bernard Vauquois. 1968. A survey of formal grammars and algorithms for recognition and transformation in machine translation. In *IFIP Congress-68*, pages 254–260, Edimburgh.

Jean Véronis. 1998. A study of polysemy judgements and inter-annotator agreement. In *Proceedings of the Senseval workshop*. Citeseer.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual e5 text embeddings: A technical report. *arXiv preprint arXiv:2402.05672*.

Yinfei Yang, Daniel Cer, Amin Ahmad, Mandy Guo, Jax Law, Noah Constant, Gustavo Hernandez Abrego, Steve Yuan, Chris Tar, Yun-hsuan Sung, Brian Strope, and Ray Kurzweil. 2020. Multilingual universal sentence encoder for semantic retrieval. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 87–94, Online. Association for Computational Linguistics.

Xin Zhang, Yanzhao Zhang, Dingkun Long, Wen Xie, Ziqi Dai, Jialong Tang, Huan Lin, Baosong Yang, Pengjun Xie, Fei Huang, and 1 others. 2024. mgte: Generalized long-context text representation and reranking models for multilingual text retrieval. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1393–1412.

A Availability of Glosses in DBnary Translations

Table 5 gives statistics on the availability of textual glosses in translations in DBnary, by source language. These statistics are computed on the `20250320` version of DBnary.

Language	#transl	#text gloss	%text gloss
Bulgarian	30166	1393	(4.6%)
Catalan	455532	195919	(43%)
Danish	47669	6852	(14.4%)
German	1066903	854695	(80.1%)
Greek (modern)	221162	26236	(11.9%)
English	3315080	3210642	(96.8%)
Finnish	232696	232113	(99.7%)
French	1337057	462564	(34.6%)
Irish (Gaeilge)	10655	6264	(58.8%)
Serbo Croat	607	0	(0%)
Indonesian	10207	0	(0%)
Italian	162006	112117	(69.2%)
Japanese	216387	55494	(25.6%)
Kurdish	750368	73549	(9.8%)
Latin	25049	5472	(21.8%)
Lithuanian	156364	156232	(99.9%)
Malagasy	148776	0	(0%)
Dutch	311370	249509	(80.1%)
Norwegian	70061	63482	(90.6%)
Polish	687603	0	(0%)
Portuguese	312860	76335	(24.4%)
Russian	760150	317534	(41.8%)
Spanish	227402	8289	(3.6%)
Swedish	416744	335938	(80.6%)
Turkish	196182	31037	(15.8%)
Chinese	1167362	276028	(23.6%)
Total	**12336418**	**6757694**	**(54.8%)**

Table 5: Statistics on the availability of textual glosses (short form designating a lexical sense definition), compared with numeric glosses (giving the lexical sense number) and redundant glosses (glossing giving both a sense number AND a short form of a definition), along with the total number of translations, by languages.