LREC 2020 Workshop
Language Resources and Evaluation Conference
11–16 May 2020

Globalex Workshop on Linked Lexicography

PROCEEDINGS

Editors:

Ilan Kernerman, Simon Krek, John P. McCrae,

Jorge Gracia, Sina Ahmadi and Besim Kabashi

Proceedings of the LREC 2020
Globalex Workshop on Linked Lexicography

Edited by: Ilan Kernerman, Simon Krek, John P. McCrae, Jorge Gracia, Sina Ahmadi and Besim Kabashi

ISBN: 979-10-95546-46-7
EAN: 9791095546467

For more information:
European Language Resources Association (ELRA)
9 rue des Cordelières
75013, Paris
France
http://www.elra.info
Email: lrec@elda.org

Introduction to the Proceedings of Globalex 2020 Workshop on Linked Lexicography

Ilan Kernerman[1], Simon Krek[2], John P. McCrae[3],
Jorge Gracia[4], Sina Ahmadi[3] and Besim Kabashi[5]

[1] K Dictionaries, Israel
[2] Jožef Stefan Institute, Slovenia
[3] SFI Insight Research Centre for Data Analytics, Data Science Institute, National University of Ireland Galway, Ireland
[4] Aragon Institute of Engineering Research (I3A), University of Zaragoza, Spain
[5]Friedrich-Alexander University of Erlangen-Nuremberg and Ludwig-Maximilian University of Munich, Germany

ilan@kdictionaries.com, simon.krek@ijs.si, john@mccr.ae, jogracia@unizar.es,
sina.ahmadi@insight-centre.org, besim.kabashi@fau.de

Abstract

The 3rd GLOBALEX Workshop at LREC 2020 has the focus of linking data from different lexicographic resources, highlighting aspects related to the automated linking of content from dictionaries and other lexical sources, with the aim of linguistic data enrichment and reinforcement. The main track of the workshop includes general research papers and is supplemented by two specific tracks, on linking monolingual data and linking bilingual and multilingual data, respectively, each combined with a shared task. The monolingual linking task was conducted as part of the ELEXIS project and the results were evaluated against novel dictionary linking data covering 15 languages developed in this project. The bilingual and multilingual linking was conducted as part of the third edition of the Translation Inference Across Dictionaries (TIAD) shared task and covered three languages matched against language pairs of K Dictionaries. These workshop proceedings include a total of 19 papers, abstracts and system descriptions, in addition to the introduction, reporting on new methodologies and techniques applied to enhance the linking of different types of lexicographic resources.
Keywords: linked lexicography, monolingual, bilingual, multilingual

1 Preface

The third Globalex workshop in conjunction with the LREC conference series[1] has become one of the numerous casualties of the COVID-19 epidemic, since LREC 2020 including all satellite workshops had to be cancelled, but its substance may live on in these proceedings.

The workshop topic of linked lexicography is embodied in these pages in the form of its 19 would-be presentations, including full papers, extended abstracts and system descriptions by scholars from across Europe and elsewhere. The focus is on linking data from different lexicographic resources, highlighting automated processes, in the aim of linguistic data enrichment and enhancement.

Linking lexicographic data sets to each other and with other lexical resources, and the interoperability of lexicography with linguistic linked data (LLD) methodologies in particular and semantic web technologies in general, have increasingly been gaining attention in recent years, becoming a subject for research projects by and collaboration between the academia and industry, including support of

[1] https://globalex2020.globalex.link//

the public sector. Most notably, the W3C community group on Ontology-Lexica[2] was established following the release of the lemon model, which constituted the first de-facto standard for representing ontology-lexica, with the mission to "develop models for the representation of lexica (and machine readable dictionaries) relative to ontologies" [17]. The ensuing OntoLex-lemon model [19] has served since 2016 as the leading option for converting lexicographic data into LLD, and was updated and finetuned through the Lexicog module released in 2019[3]. This trend has been complemented since 2015 by relevant literature (e.g. [10, 14, 4]), conference papers (e.g. [7, 1, 5, 12, 6]) and mainly EU-funded projects (LDL4HELTA, ELEXIS, Prêt-à-LLOD), and continues to be pursued and advanced as further attested in this volume.

The main track of the workshop included general research papers on linked lexicography and related topics, described in section 2. This was complemented by two in-focus tracks with corresponding shared tasks, on linking monolingual lexicographic resources, in conjunction with ELEXIS, described in section 3, and on linking bilingual and multilingual lexicographic resources, in conjunction with TIAD shared task (TIAD 2020), described in section 4. In section 5 we report on our conclusions.

Globalex 2020 was endorsed by Globalex[4], the Global Alliance for Lexicography, following up on the first Globalex Workshop on Lexicographic Resources and Human Language Technology at LREC 2016[5] and the second Globalex Workshop on Lexicography and Wordnets at LREC 2018[6], with the support of ELEXIS and TIAD.

2 Linking Lexicography

This general track of the workshop includes three papers and three abstracts directly related to the workshop's main theme of linked lexicography as well as three papers and two abstracts on other lexicographic topics.

The first paper, **Modelling frequency and attestations for OntoLex-Lemon**, by Christian Chiarcos, Maxim Ionov, Jesse de Does, Katrien Depuydt, Anas Fahad Khan, Sander Stolk, Thierry Declerck and John Philip McCrae, describes the new FrAC extension of the OntoLex model for corpus-related information. The Ontolex-Lemon W3C community has been shaping up since 2012 and released the state-of-the-art Lexicog module for lexicography in 2019. FrAC aims to make new grounds dealing with "supplementary information drawn from corpora such as frequency information, links to attestations in corpora, and collocation data ... that not only covers the requirements of digital lexicography, but also accommodates essential data structures for lexical information in natural language processing". The paper also illustrates use-cases that implement the model on diverse resources serving different purposes.

The next paper, **SynSemClass linked lexicon: Mapping synonymy between languages**, by Zdenka Uresova, Eva Fucikova, Eva Hajicova and Jan Hajic, presents a cross-lingual study of verb synonymy through verb classes, valency information and semantic roles and "reports on an extended version of a synonym verb class lexicon ... [which] stores cross-lingual semantically similar verb senses in synonym classes extracted from a richly annotated parallel corpus", making use of valency relations and linking them to semantic roles and external lexicons. The aims include comparing "semantic roles and their syntactic properties and features across languages within and across synonym groups, [offering] gold standard data for automatic NLP experiments with such synonyms" and, most notably, building "an

[2] https://www.w3.org/community/ontolex/
[3] https://www.w3.org/2019/09/lexicog/
[4] https://globalex.link/
[5] https://globalex2016.globalex.link/
[6] https://globalex2018.globalex.link/

event type ontology that can be referenced and used as a human-readable and human-understandable "database" for all types of events, processes and states". In addition to describing its content, the authors present a preliminary design of a linked data-compatible format of their lexicon.

The third paper, **Representing etymology in the LiLa knowledge base of linguistic resources for Latin**, by Francesco Mambrini and Marco Passarotti, describes "the process of inclusion of etymological information in a knowledge base of interoperable Latin linguistic resources", applying Linked Open Data principles based on the Ontolex-Lemon ontology and the lemonEty extension. The authors present their motivation, methodology and modelling strategies as well as possible applications and further developments.

The papers are followed by three abstracts. The first, **An automatically generated Danish Renaissance Dictionary**, by Mette-Marie Møller Svendsen, Nicolai Hartvig Sørensen and Thomas Troelsgård, describes "[b]uilding a period dictionary by reducing and merging relevant existing dictionary resources". The main goal of this project is "to present a series of Danish hymn books from the Lutheran Reformation" including digitizing and making searchable texts and music as well as access to partially digitized dictionaries that are relevant to this period, including an integrated dictionary function to look up words in the text and present sense keywords extended from the dictionary entries as well links to full dictionary entries."

The second abstract, **Linking the Open Dutch Wordnet with Dutch lexicographic resources**, by Thierry Declerck, describes ongoing work on linking wordnet resources from the Open Multilingual Wordnet initiative to morphological ones, with the aim of mutual enrichment. At the first stage, Romance language resources were mapped onto the OntoLex-Lemon model, with interlinking carried out "automatically ... by selecting the identical lemmas encoded on both sides, based on string matching [followed by m]anual correction for linking homographs to their ... Wordnet entries"; as a result, morphological variants were added to the lexical data, realizing the Wordnet concepts, with the added value of formulating lexical restrictions. The experiment continued with interlinking Wordnets to richer resources (beyond solely morphology) in the form of a comprehensive dictionary of Dutch, which turned out to be more complex and required metadata comparisons.

The third abstract, **Widening the discussion on 'false friends' in multilingual dictionaries and linked lexicographic resources**, by Hugo Gonçalo Oliveira and Ana Luís, discusses potential problems of false friends in the multilingual alignment of existing wordnets, with a specific use case providing examples of erroneous alignments between English and Portuguese synsets. The authors suggest to "exploit lists of false friends from the literature for cleaning multilingual wordnets, ... remove false friends from linked synsets, or even to remove the connections between those synsets,... [and that] an RDF property could perhaps be used for explicitly linking pairs of lexical items, in different languages ...".

The second part of this general track section includes three papers relating to specific languages and two abstracts on domain-specific lexicography/terminology. The first paper, **Pinchah Kristang: A dictionary of Kristang**, by Luís Morgado da Costa, describes "the development and current state of ... an online dictionary for Kristang[,] a critically endangered language of the Portuguese-Eurasian communities residing mainly in Malacca and Singapore". This dictionary constitutes a central tool to the revitalization of the language, collating "information from multiple sources, including existing dictionaries and wordlists, ongoing language documentation work, and new words that emerge regularly from relexification efforts by the community", and is powered by the Princeton and Open Kristang wordnets.

The next two papers concern Scandinavian languages from the opposite "privileged" extreme of

the scale. **Building sense representations in Danish by combining word embeddings with lexical resources**, by Ida Rørmann Olsen, Bolette Pedersen and Asad Sayeed, concerns a project for identification of suitable sense representation for NLP in Danish. The authors "investigate sense inventories that correlate with human interpretations of word meaning and ambiguity as typically described in dictionaries and wordnets and that are well reflected distributionally as expressed in word embeddings ... study a number of highly ambiguous Danish nouns and examine the effectiveness of sense representations constructed by combining vectors from a distributional model with the information from a wordnet. We establish representations based on centroids obtained from wordnet synests and example sentences as well as representations established via a clustering approach [and] tested in a word sense disambiguation task[, concluding] that the more information extracted from the wordnet entries ... the more successful the sense representation vector".

Then, **Towards a Swedish Roget-style thesaurus for NLP**, by Niklas Zechner and Lars Borin, examines whether and how a digitized Swedish thesaurus originally published in 1930 can serve multiple NLP applications, concluding that "to be useful in our NLP systems, polysemous lexical items need to be disambiguated, and a large amount of modern vocabulary must be added in the proper places". The authors describe "experiments aiming at automating these two tasks, at least in part, where we use the structure of an existing Swedish semantic lexicon" both for disambiguating ambiguous thesaurus entries and adding new entries.

The abstract, **Design and development of an adaptive web application for OLIVATERM**, by Mercedes Roldán Vendrell, describes the project dedicated to designing "the first systematic multilingual terminological dictionary in the scientific and socio-economic area of the olive grove and olive oils", and the work that continues on the development of "a multichannel technological solution [to enable] greater and more efficient transfer to the business sector" combined with a responsive website and an interactive web-based application offering dynamic transfer of relevant information to and from users.

In the last abstract concluding this section, **Building a domain-specific bilingual lexicon resource with Sketch Engine and Lexonomy: Taking ownership of the issues**, Zaida Bartolomé-Díaz and Francesca Frontini question the value of modern methods to accelerate and standardize the elaboration of specialized bilingual dictionaries, "offering not only a relation of terms, but also a representation of a conceptual field" in contrast to "the viability of their use by a lambda user and the previous knowledge" needed for such efficient use, and the possible problems that might occur. The authors propose methodological solutions based on a small corpus consisting of 82 documents extracted from the web, using a list of selected terms, aimed to create automatically a dictionary extract of about 25 terms.

3 Linking monolingual lexicographic resources

3.1 Task Description

The Monolingual Word Sense Alignment (MWSA) task was concerned with the linking of two dictionaries in a single language at the sense level. For example, multiple senses of a word such as for "chair", the sense with definition "a seat for one person, with a support for the back" would be linked to another sense in another dictionary "a movable single seat with a back", while the sense for "the officer who presides at the meetings of an organization" would be linked to "the presiding officer of an assembly". The dataset used for this evaluation was the one prepared by [2] which covers 15 languages and includes alignments between 17 dictionaries. This resource lists all the sense links between the two dictionaries classified with one of the following relationships:

Language	Metric	Baseline	ACDH	RACAI	UNIOR NLP
English	Accuracy	0.752	0.763	**0.798**	0.759
	Precision	0.000	0.619	**0.746**	0.586
	Recall	0.000	0.782	0.353	**0.692**
	F-Measure	0.000	**0.691**	0.480	0.634
Basque	Accuracy	**0.789**	0.407	-	-
	Precision	0.211	**0.223**	-	-
	Recall	0.050	**0.738**	-	-
	F-Measure	0.081	**0.342**	-	-
Bulgarian	Accuracy	**0.728**	0.395	-	-
	Precision	0.250	**0.331**	-	-
	Recall	0.011	**0.842**	-	-
	F-Measure	0.020	**0.475**	-	-
Danish	Accuracy	**0.817**	0.522	-	-
	Precision	**0.300**	0.253	-	-
	Recall	0.023	**0.756**	-	-
	F-Measure	0.043	**0.379**	-	-
Dutch	Accuracy	0.936	0.940	**0.944**	0.931
	Precision	0.000	0.636	**0.846**	0.455
	Recall	0.000	**0.241**	0.190	0.086
	F-Measure	0.000	**0.350**	0.310	0.145
Estonian	Accuracy	0.482	**0.565**	-	-
	Precision	0.545	**0.707**	-	-
	Recall	0.093	**0.806**	-	-
	F-Measure	0.159	**0.754**	-	-
German	Accuracy	0.7777	**0.798**	-	-
	Precision	0.000	**0.738**	-	-
	Recall	0.000	**0.608**	-	-
	F-Measure	0.000	**0.667**	-	-
Hungarian	Accuracy	**0.940**	-	-	-
	Precision	**0.053**	-	-	-
	Recall	**0.012**	-	-	-
	F-Measure	**0.020**	-	-	-
Irish	Accuracy	**0.583**	0.549	-	-
	Precision	**0.680**	0.631	-	-
	Recall	0.185	**0.891**	-	-
	F-Measure	0.291	**0.739**	-	-
Italian	Accuracy	0.693	0.537	0.761	**0.766**
	Precision	0.000	0.418	**0.760**	0.729
	Recall	0.000	0.719	0.333	**0.754**
	F-Measure	0.000	0.529	0.463	**0.741**
Portuguese	Accuracy	**0.921**	0.870	-	0.933
	Precision	0.083	0.311	-	**0.541**
	Recall	0.024	0.762	-	**0.786**
	F-Measure	0.037	0.441	-	**0.641**
Russian	Accuracy	**0.754**	0.606	-	-
	Precision	**0.438**	0.372	-	-
	Recall	0.179	**0.821**	-	-
	F-Measure	0.255	**0.512**	-	-
Serbian	Accuracy	0.853	**0.599**	-	-
	Precision	0.000	**0.190**	-	-
	Recall	0.000	**0.464**	-	-
	F-Measure	0.000	**0.269**	-	-
Slovene	Accuracy	**0.834**	0.442	-	-
	Precision	0.100	**0.173**	-	-
	Recall	0.009	**0.587**	-	-
	F-Measure	0.017	**0.268**	-	-
Spanish	Accuracy	0.678	-	0.786	**0.829**
	Precision	0.255	-	0.667	**0.742**
	Recall	0.127	-	0.655	**0.891**
	F-Measure	0.170	-	0.661	**0.810**
Average	Accuracy	0.769	0.615	0.822	**0.844**
	Precision	0.194	0.431	**0.755**	0.611
	Recall	0.048	**0.694**	0.383	0.642
	F-Measure	0.074	0.494	0.478	**0.594**

Table 1: Results of the evaluation of the MWSA task by team and language

Exact The sense are the same, for example the definitions are simply paraphrases

Broader The sense in the first dictionary completely covers the meaning of the sense in the second dictionary and is applicable to further meanings

Narrower The sense in the first dictionary is entirely covered by the sense of the second dictionary, which is applicable to further meanings

Related There are cases when the senses may be equal but the definitions in both dictionaries differ in key aspects

None There is no match for this sense

The evaluation of the shared task therefore used multiple metrics to evaluate the results of the system. Firstly, *accuracy* measured the total number of links for which the correct class of relationship was predicted. Secondly, we provided *recall*, *precision* and *F-Measure* scores based on a 2-class classification problem, where the 'exact', 'broader', 'narrower' and 'related' links were merged into a single positive class. This was motivated by the fact that many applications do not care about the specific type of link and that detecting the presence of the link was harder task from predicting the type of the link. We provided this analysis for each of the languages and scored the systems overall based on a macro-average of the accuracy, precision, recall and F-Measure.

3.2 Participants

The task was organized using CodaLab[7] and three external teams[8] participated, although not all teams participated for all languages. The baseline model was quite simple: for each sense pair the Jaccard similarity of the gloss was calculated, then the Hungarian Algorithm [15] was used to find the most likely unique assignment between these senses. The baseline only predicted the 'exact' class (and 'none') so it was expected that the results would be quite poor. The other approaches taken by participants were as follows:

RACAI The RACAI system viewed this task as a case of word-sense disambiguation, from this multiple features were extracted including scores based on the Lesk algorithm [16] as well as features from BERT [8] and other features, which were combined using a random forest [13].

ACDH A variety of features were combined in this approach including simple similarity methods such as used in the baseline as well as similarities coming from ELMo [22] and BERT. These were also combined using a supervised learning framework, and different settings were used for each language.

UNIOR NLP This approach used BERT as well as Siamese LSTMs [21] improved with lexico-semantic information related to the lemma's part-of-speech category.

The overall results are presented in Table 1, and we can see that the overall strongest result in accuracy and F-Measure was obtained by the UNIOR NLP team. However, all systems can be said to have performed best on some of the tasks (even the baseline) and given that all systems used BERT, more research is needed into the best way to fine-tune BERT for this task.

[7] `https://competitions.codalab.org/competitions/22163`

[8] A fourth team participated, but withdrew after submitting results

4 Linking bilingual and multilingual lexicographic resources

In this section we give an overview of the goals and results of the 3rd edition of the Translation Inference Across Dictionaries (TIAD) initiative, co-located with Globalex 2020.

4.1 Task description

The shared task for Translation Inference Across Dictionaries was aimed at exploring methods and techniques for automatically generating new bilingual (and multilingual) dictionaries from existing ones. The main aim of TIAD is to support a coherent experiment framework that enables reliable validation of results and solid comparison of the processes used. This initiative also aims to enhance further research on the topic of inferring translations across languages, and continues the first and second TIAD workshops, which took place on June 18, 2017 in Galway (Ireland) and in Leipzig (Germany) on May 20, 2019, respectively, co-located with the 1st and 2nd editions of the Language Data and Knowledge (LDK) conference.

The experimental setup for this year's evaluation campaign has been the same as in the 2nd TIAD edition [11] with minor differences such as the inclusion of a validation data set (a sample of 5% of the test data set) and the curation of the test data (see later). The participating systems were asked to generate new translations automatically among three languages - English, French, Portuguese - based on known translations contained in the Apertium RDF graph[9]. As these languages (EN, FR, PT) are not directly connected in this graph, no translations can be obtained directly among them there. Based on the available RDF data, the participants had to apply their methodologies to derive translations, mediated by any other language in the graph, between the pairs EN/FR, FR/PT and PT/EN. See the TIAD 2020 website[10] for more technical details on the experimental setup and results.

The evaluation of the results was carried out by the organisers against manually compiled pairs of K Dictionaries (KD), extracted from its Global Series (https://lexicala.com/).

4.2 Results

Nine systems participated in the shared task, coming from four different teams. The participant teams submitted a system description paper including: a description of their system or systems, the way data was processed, the applied algorithms, the obtained results, as well as the conclusions and ideas for future improvements. The system papers were reviewed by the organising committee to confirm that all these aspects were well covered.

This is the list of the participating teams along with a short description of their contributions:

CUD. A *multi-strategy* system was deveoped by Centro Univesritario de la Defensa (CUD), Spain, which combines several strategies to analyse the Apertium RDF graph, taking advantage of characteristics such as translation using multiple paths, synonyms and similarities between lexical entries from different lexicons and cardinality of possible translations through the graph. Several combinations of such strategies were presented to the shared task, showing that the combination of all of them produces better results than without joining all the strategies.

[9] `http://linguistic.linkeddata.es/apertium/`
[10] `https://tiad2020.unizar.es`

NUIG. This is the contribution of National University of Ireland Galway (NUIG) to TIAD. The proposed system combines unsupervised NLP and Graph Metrics for Translation Inference. This system includes graph-based metrics calculated using novel algorithms, with an unsupervised document embedding tool called ONETA and an unsupervised multi-way neural machine translation method. The results improve the system that the authors presented in the last TIAD edition [18] and produces the highest precision among all systems in the task while preserving a reasonable recall.

ACoLi. The Applied Computational Linguistics (ACoLi), Goethe University Frankfurt, Germany, contributed with a method based on symbolic methods and the propagation of concepts over a graph of interconnected dictionaries, which evolves the system presented by the authors in the previous TIAD edition [9]. Given a mapping from source language words to lexical concepts (e.g., synsets) as a seed, they use bilingual dictionaries to extrapolate a mapping of pivot and target language words to these lexical concepts. Translation inference is then performed by looking up the lexical concept(s) of a source language word and returning the target language word(s) for which these lexical concepts have the respective highest score. They participated with two instantiations of such a system: one using WordNet synsets as concepts, and one using lexical entries (translations) as concepts.

UNIZAR. University of Zaragoza (UNIZAR), Spain, contributed with two different systems to the shared task. On the one hand `Cycles-OTIC`, a hybrid technique based on graph exploration that combines a method that explores the density of cycles in the translations graph [24] with the translations obtained by the One Time Inverse Consultation (OTIC) method [23], which obtained better coverage than OTIC alone but slightly reduced precision. On the other hand, *Cross-lingual embeddings*, based on the distribution of embeddings across languages [3], were used to build cross-lingual word embeddings trained with monolingual corpora and mapped afterwards through an intermediate language.

We have run two baselines to be compared with the participating systems:

Baseline 1 - Word2Vec. The method uses Word2Vec [20] to transform the graph into a vector space. A graph edge is interpreted as a sentence and the nodes are word forms with their POS tag. Word2Vec iterates multiple times over the graph and learns multilingual embeddings (without additional data). For a given input word, we calculated a distance based on the cosine similarity of a word to every other word with the target-POS tag in the target language. In our evaluation, we applied an arbitrary threshold of 0.5 to the confidence degree. Note that in the TIAD 2020 edition the Word2Vec baseline, although based on the same principles of TIAD 2019, has been re-implemented and re-trained and lead to different results than in the previous TIAD edition.

Baseline 2 - OTIC. The idea of the One Time Inverse Consultation (OTIC) method [23] is to explore, for a given word, the possible candidate translations that can be obtained through intermediate translations in the pivot language. Then, a score is assigned to each candidate translation based on the degree of overlap between the pivot translations shared by both the source and target words. In our evaluation, we have applied the OTIC method using Spanish as pivot language, and using an arbitrary threshold of 0.5.

The results can be seen in Table 2 and demonstrate that most of the systems show good precision (all of them over 0.6 but the Word2Vec baseline) but a lesser recall (none of them reached 0.5). The OTIC baseline continues being a simple but hard to beat baseline. Overall the results have been better that the ones obtained in TIAD 2019 [11], with F-measure results in the range [0.25, 0.56], compared with the range [0.11, 0.37] in 2019. One of the main reasons, in addition the particular systems improvements, is that the golden standard data have been curated with respect to the previous version in two aspects:

by removing duplicated entries caused by the presence of non-breaking spaces in Apertium, and by removing some entries that were not in the intersection between Apertium and KD data; thus leading to an increased recall.

System	Precision	Recall	F-measure	Coverage
BASELINE(OTIC)	0.7	0.47	0.56	0.7
Cycles-OTIC	0.64	0.47	0.54	0.76
NUIG	0.77	0.35	0.49	0.54
Multi-StrategyI+II+III+IV	0.61	0.33	0.43	0.63
Multi-StrategyI+II+III	0.62	0.33	0.43	0.63
CL-embeddings	0.62	0.32	0.42	0.59
Multi-StrategyI+II	0.65	0.3	0.4	0.59
ACOLIbaseline	0.6	0.28	0.38	0.48
BASELINE(Word2Vec)	0.3	0.37	0.33	0.68
Multi-StrategyI	0.63	0.22	0.32	0.44
ACOLIwordnet	0.61	0.16	0.25	0.28

Table 2: TIAD 2020 averaged system results, ordered by F-measure in descending order.

5 Conclusion

While this workshop has not been able to physically take place this year, these proceedings show that the work in the area of digital lexicography is still very much alive. In particular, with the introduction of the two shared tasks, we have made a closer connection between lexicographers and computer scientists, allowing state-of-the-art methods in natural language processing including deep learning to be applied to solve challenges in lexicography. Moreover, we continue to see the value in semantic web technologies for the representation of lexicographic resources and are encouraged to see more work supporting this and the use of linked data methodologies in lexicography in line with the workshop's theme of *linked lexicography*.

References

[1] Frank Abromeit, Christian Chiarcos, Christian Fäth, and Maxim Ionov. Linking the Tower of Babel: modelling a massive set of etymological dictionaries as RDF. In *Proceedings of the 5th Workshop on Linked Data in Linguistics (LDL-2016): Managing, Building and Using Linked Language Resources*, pages 11–19, 2016.

[2] Sina Ahmadi, John P. McCrae, Sanni Nimb, Thomas Troelsgård, Sussi Olsen, Bolette S. Pedersen, Thierry Declerck, Tanja Wissik, Monica Monachini, Andrea Bellandi, Fahad Khan, Irene Pisani, Simon Krek, Veronika Lipp, Tamás Váradi, László Simon, András Győrffy, Carole Tiberius, Tanneke Schoonheim, Yifat Ben Moshe, Maya Rudich, Raya Abu Ahmad, Dorielle Lonke, Kira Kovalenko, Margit Langemets, Jelena Kallas, Oksana Dereza, Theodorus Fransen, David Cillessen, David Lindemann, Mikel Alonso, Ana Salgado, José Luis Sancho, Rafael-J. Ure na Ruiz, Kiril Simov, Petya Osenova, Zara Kancheva, Ivaylo Radev, Ranka Stanković, Cvetana Krstev, Biljana Lazić, Aleksandra Marković, Andrej Perdih, and Dejan Gabrovšek. A Multilingual Evaluation Dataset for Monolingual Word Sense Alignment. In *Proceedings of the 12th Language Resource and Evaluation Conference (LREC 2020)*, 2020.

[3] Mikel Artetxe, Gorka Labaka, and Eneko Agirre. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 789–798, 2018.

[4] Julia Bosque-Gil, Jorge Gracia, and Asunción Gómez-Pérez. Linked data in lexicography. *Kernerman Dictionary News*, (24):19–24, 2016.

[5] Julia Bosque-Gil, Jorge Gracia, and Elena Montiel-Ponsoda. Towards a Module for Lexicography in OntoLex. In *LDK Workshops*, pages 74–84, 2017.

[6] Julia Bosque-Gil, Dorielle Lonke, Jorge Gracia, and Ilan Kernerman. Validating the OntoLex-lemon Lexicography Module with K Dictionaries' Multilingual Data. In *Electronic lexicography in the 21st century. Proceedings of the eLex 2019 conference*, pages 726–746, 2019.

[7] Thierry Declerck, Eveline Wandl-Vogt, and Karlheinz Mörth. Towards a Pan-European lexicography by means of linked (open) data. *Proceedings of eLex*, pages 342–355, 2015.

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[9] Kathrin Donandt and Christian Chiarcos. Translation inference through multi-lingual word embedding similarity. In *Proceedings of TIAD-2019 Shared Task – Translation Inference Across Dictionaries, at 2nd Language Data and Knowledge (LDK) conference*. CEUR-WS, May 2019.

[10] Jorge Gracia. Multilingual dictionaries and the Web of Data. *Kernerman Dictionary News*, (23):1–4, 2015.

[11] Jorge Gracia, Besim Kabashi, Ilan Kernerman, Marta Lanau-Coronas, and Dorielle Lonke. Results of the Translation Inference Across Dictionaries 2019 Shared Task. In Jorge Gracia, Besim Kabashi, and Ilan Kernerman, editors, *Proceedings of TIAD-2019 Shared Task – Translation Inference Across Dictionaries co-located with the 2nd Language, Data and Knowledge Conference (LDK 2019)*, pages 1–12, Leipzig (Germany), 2019. CEUR Press.

[12] Jorge Gracia, Ilan Kernerman, and Julia Bosque-Gil. Toward linked data-native dictionaries. In *Electronic Lexicography in the 21st Century: Lexicography from Scratch. Proceedings of the eLex 2017 conference*, pages 19–21, 2017.

[13] Tin Kam Ho. Random decision forests. In *Proceedings of the 3rd international conference on document analysis and recognition*, volume 1, pages 278–282. IEEE, 1995.

[14] Bettina Klimek and Martin Brümmer. Enhancing lexicography with semantic language databases. *Kernerman Dictionary News*, (23):5–10, 2015.

[15] Harold W Kuhn. The Hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955.

[16] Michael Lesk. Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. In *Proceedings of the 5th annual international conference on Systems documentation*, pages 24–26, 1986.

[17] John McCrae, Guadalupe Aguado de Cea, Paul Buitelaar, Philipp Cimiano, Thierry Declerck, Asunción Gómez-Pérez, Jorge Gracia, Laura Hollink, Elena Montiel-Ponsoda, Dennis Spohr, and Tobias Wunner. Interchanging lexical resources on the Semantic Web. *Language Resources and Evaluation*, 46(6):701–709, 2012.

[18] John P. McCrae. TIAD shared task 2019: Orthonormal explicit topic analysis for translation inference across dictionaries. In *Proceedings of TIAD-2019 Shared Task – Translation Inference Across Dictionaries, at 2nd Language Data and Knowledge (LDK) conference*. CEUR-WS, May 2019.

[19] John P. McCrae, Julia Bosque-Gil, Jorge Gracia, Paul Buitelaar, and Philipp Cimiano. The OntoLex-Lemon Model: development and applications. In *Proceedings of eLex 2017*, pages 587–597, 2017.

[20] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2013.

[21] Jonas Mueller and Aditya Thyagarajan. Siamese recurrent architectures for learning sentence similarity. In *thirtieth AAAI conference on Artificial Intelligence*, 2016.

[22] Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*, 2018.

[23] Kumiko Tanaka and Kyoji Umemura. Construction of a Bilingual Dictionary Intermediated by a Third Language. In *COLING*, pages 297–303, 1994.

[24] Marta Villegas, Maite Melero, Núria Bel, Jorge Gracia, and Núria Bel. Leveraging RDF Graphs for Crossing Multiple Bilingual Dictionaries. In *Proceedings of trhe 10th Language Resources and Evaluation Conference (LREC'16) Portorož (Slovenia)*, pages 868–876, Paris, France, may 2016. European Language Resources Association (ELRA).

Organizers

Ilan Kernerman, K Dictionaries
Simon Krek, Globalex, Jožef Stefan Institute

Track 1 Organizers

John McCrae, National University of Ireland Galway
Sina Ahmadi, National University of Ireland Galway

Track 2 Organizers

Jorge Gracia, University of Zaragoza
Besim Kabashi, Friedrich-Alexander University of Erlangen-Nuremberg and Ludwig-Maximilian University of Munich

Program Committee

Anna Braasch. University of Copenhagen, Denmark
Sara Carvalho. University of Aveiro, Portugal
Philip Cimiano. University of Bielefeld, Germany
Rute Costa, Universidade Nova de Lisboa, Portugal
Thierry Fontenelle. European Investment Bank, Luxembourg
Radovan Garabik. Ľ. Štúr Institute of Linguistics, Slovakia
Jorge Gracia, University of Zaragoza, Spain
Dagmar Gromann. University of Vienna, Austria
Ales Horak. Masaryk University, Czech Republic
Besim Kabashi. Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany
Ilan Kernerman. K Dictionaries, Israel
Iztok Kosem. Jožef Stefan Institute, Slovenia
Simon Krek. Jožef Stefan Institute, Slovenia
Nikola Ljubešić. Jožef Stefan Institute, Slovenia
Dorielle Lonke. K Dictionaries, Israel
Patricia Martín Chozas. Madrid Polytechnic University, Spain
John Philip McCrae. National University of Ireland Galway, Ireland
Krzysztof Nowak. Institute of Polish Language, Poland
Maciej Piaceski. Wroclaw University of Science and Technology, Poland
Carole Tiberius. Instituut voor Nederlandse Lexicologie, Netherlands
Lars Trap-Jensen. Society for Danish Language and Literature, Denmark
Marieke van Erp. KNAW Humanities Cluster, Netherlands

Table of Contents

Modelling Frequency and Attestations for OntoLex-Lemon
Christian Chiarcos, Maxim Ionov, Jesse de Does, Katrien Depuydt, Anas Fahad Khan, Sander Stolk, Thierry Declerck and John Philip McCrae 1

SynSemClass Linked Lexicon: Mapping Synonymy between Languages
Zdenka Uresova, Eva Fucikova, Eva Hajicova and Jan Hajic 10

Representing Etymology in the LiLa Knowledge Base of Linguistic Resources for Latin
Francesco Mambrini and Marco Passarotti 20

An automatically generated Danish Renaissance Dictionary
Mette-Marie Møller Svendsen, Nicolai Hartvig Sørensen and Thomas Troelsgård 29

Towards an Extension of the Linking of the Open Dutch WordNet with Dutch Lexicographic Resources
Thierry Declerck 33

Widening the Discussion on "False Friends" in Multilingual Wordnets
Hugo Gonçalo Oliveira and Ana Luís 36

Pinchah Kristang: A Dictionary of Kristang
Luís Morgado da Costa 37

Building Sense Representations in Danish by Combining Word Embeddings with Lexical Resources
Ida Rørmann Olsen, Bolette Pedersen and Asad Sayeed 45

Towards a Swedish Roget-Style Thesaurus for NLP
Niklas Zechner and Lars Borin 53

Design and development of an adaptive web application for OLIVATERM
Mercedes Roldán Vendrell 61

Building a domain-specific bilingual lexicon resource with Sketchengine and Lexonomy: Taking Ownership of the Issues
Zaida Bartolomé-Díaz and Francesca Frontini 62

MWSA Task at GlobaLex 2020: RACAI's Word Sense Alignment System using a Similarity Measurement of Dictionary Definitions
Vasile Pais, Dan Tufiș and Radu Ion 69

UNIOR NLP at MWSA Task - GlobaLex 2020: Siamese LSTM with Attention for Word Sense Alignment
Raffaele Manna, Giulia Speranza, Maria Pia di Buono and Johanna Monti 76

Implementation of Supervised Training Approaches for Monolingual Word Sense Alignment: ACDH-CH System Description for the MWSA Shared Task at GlobaLex 2020
Lenka Bajcetic and Seung-bin Yim 84

NUIG at TIAD: Combining Unsupervised NLP and Graph Metrics for Translation Inference
John Philip McCrae and Mihael Arcan 92

Translation Inference by Concept Propagation
Christian Chiarcos, Niko Schenk and Christian Fäth 98

Graph Exploration and Cross-lingual Word Embeddings for Translation Inference Across Dictionaries
Marta Lanau-Coronas and Jorge Gracia 106

Multi-Strategy system for translation inference across dictionaries
Lacramioara Dranca 111

Conference Program

GlobalLex Presentations

Main Track

Modelling Frequency and Attestations for OntoLex-Lemon
Christian Chiarcos, Maxim Ionov, Jesse de Does, Katrien Depuydt, Anas Fahad Khan, Sander Stolk, Thierry Declerck and John Philip McCrae

SynSemClass Linked Lexicon: Mapping Synonymy between Languages
Zdenka Uresova, Eva Fucikova, Eva Hajicova and Jan Hajic

Representing Etymology in the LiLa Knowledge Base of Linguistic Resources for Latin
Francesco Mambrini and Marco Passarotti

An automatically generated Danish Renaissance Dictionary
Mette-Marie Møller Svendsen, Nicolai Hartvig Sørensen and Thomas Troelsgård

Towards an Extension of the Linking of the Open Dutch WordNet with Dutch Lexicographic Resources
Thierry Declerck

Widening the Discussion on "False Friends" in Multilingual Wordnets
Hugo Gonçalo Oliveira and Ana Luís

Pinchah Kristang: A Dictionary of Kristang
Luís Morgado da Costa

Building Sense Representations in Danish by Combining Word Embeddings with Lexical Resources
Ida Rørmann Olsen, Bolette Pedersen and Asad Sayeed

Towards a Swedish Roget-Style Thesaurus for NLP
Niklas Zechner and Lars Borin

Design and development of an adaptive web application for OLIVATERM
Mercedes Roldán Vendrell

GlobalLex Presentations (continued)

Building a domain-specific bilingual lexicon resource with Sketchengine and Lexonomy: Taking Ownership of the Issues
Zaida Bartolomé-Díaz and Francesca Frontini

MWSA Shared Task

MWSA Task at GlobaLex 2020: RACAI's Word Sense Alignment System using a Similarity Measurement of Dictionary Definitions
Vasile Pais, Dan Tufiș and Radu Ion

UNIOR NLP at MWSA Task - GlobaLex 2020: Siamese LSTM with Attention for Word Sense Alignment
Raffaele Manna, Giulia Speranza, Maria Pia di Buono and Johanna Monti

Implementation of Supervised Training Approaches for Monolingual Word Sense Alignment: ACDH-CH System Description for the MWSA Shared Task at GlobaLex 2020
Lenka Bajcetic and Seung-bin Yim

TIAD Shared Task

NUIG at TIAD: Combining Unsupervised NLP and Graph Metrics for Translation Inference
John Philip McCrae and Mihael Arcan

Translation Inference by Concept Propagation
Christian Chiarcos, Niko Schenk and Christian Fäth

Graph Exploration and Cross-lingual Word Embeddings for Translation Inference Across Dictionaries
Marta Lanau-Coronas and Jorge Gracia

Multi-Strategy system for translation inference across dictionaries
Lacramioara Dranca