Proceedings of the Globalex Workshop on Linked Lexicography, pages 98–105
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

Translation Inference by Concept Propagation

Christian Chiarcos, Niko Schenk, Christian Fäth
Applied Computational Linguistics (ACoLi)
Goethe University Frankfurt, Germany
{chiarcos, n.schenk, faeth}@em.uni-frankfurt.de

Abstract

This paper describes our contribution to the Third Shared Task on Translation Inference across Dictionaries (TIAD-2020). We describe an approach on translation inference based on symbolic methods, the propagation of concepts over a graph of interconnected dictionaries: Given a mapping from source language words to lexical concepts (e.g., synsets) as a seed, we 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. We present two instantiations of this system: One using WordNet synsets as concepts, and one using lexical entries (translations) as concepts. With a threshold of 0, the latter configuration is the second among participant systems in terms of F1 score. We also describe additional evaluation experiments on Apertium data, a comparison with an earlier approach based on embedding projection, and an approach for constrained projection that outperforms the TIAD-2020 vanilla system by a large margin.

Keywords: Translation Inference, Bilingual Dictionaries, Auto-Generating Dictionaries

1. Background

The Third Shared Task on Translation Inference across Dictionaries[1] (TIAD-2020) has been conducted in conjunction with the GlobaLex workshop at the 12th Language Resources and Evaluation Conference (LREC-2020, Marseille, France). As in previous editions, the objective is to automatically obtain new bilingual dictionaries based on existing ones. The evaluation is performed against a blind test set provided by a commercial partner (KDictionaries, Tel Aviv), so that a particular challenge is to optimize against data with unknown characteristics. For this edition, participants were provided with a test data excerpt to study its characteristics before the submission of final results.

Our system from TIAD-2019 (Donandt and Chiarcos, 2019) was based on translation inference over bilingual dictionaries by means of *embedding propagation*: Given word embeddings[2] for a particular pivot (seed) language, we extrapolated the embeddings of all languages with translations into (or from) the pivot language (first generation languages) by adding the corresponding English language scores. Where translations into (or from) the seed language are missing, translations into (or from) first generation languages are used to deduce embeddings in the same manner for second generation languages, etc. In this way, word embeddings are propagated through the entire dictionary graph. For predicting translations, we then use cosine similarity between source and target language vectors, constrained by a similarity threshold.

This approach is both simple and knowledge-poor; it uses no other multilingual resources than the ones provided by the task organizers, and it is particularly well-suited to address under-resourced languages as addressed in language documentation (i.e., languages for which no substantial corpora [neither monolingual, nor parallel] are available, but the majority of published language data is represented in secondary resources such as word lists and grammars). The objective for developing this system was to facilitate language contact studies (esp., cognate detection) for the Caucasus, and to address semantic similarity for low-resource language varieties for which only word lists are provided and where no basis for inducing native word or parallel embeddings exists. Despite its simplicity and the highly specialized domain of application it was designed for, our system performed well among participant systems, being a high precision system with top F1 score (precision 0.64, recall 0.22, F1 0.32; the closest competitor system had the scores: precision 0.36, recall 0.31, F1 0.32). At the same time, none of the participant systems outperformed the organizer's baselines, and we suspected the following reasons:

- The characteristics of the training data (Apertium dictionaries were designed for machine translation, i.e., to give the most *common* translation) and the evaluation data (KDictionaries dictionaries were designed for language learning, i.e., to give the most *precise* translation) may be so different that optimization against external resources (or the provided training data) does not improve performance over the evaluation data.
- Participant systems, in particular those based on word embeddings (by embedding projection or other techniques), are probably effective at capturing the main sense of a word, but they lose on secondary senses because these are under-represented in the corpora used to derive the original embeddings but over-represented in the evaluation data.

Whereas TIAD-2019 was reductionist in that every lexical entry was represented by a single vector (capturing its main sense, resp. a weighted average of all its senses), our current TIAD-2020 contribution is tailored towards the second aspect: we aim to preserve the diversity of translations provided in the training data by *propagating lexical concepts* rather than embeddings.

[1] `https://tiad2020.unizar.es/`

[2] We used 50-dimensional GloVe embeddings (Pennington et al., 2014) for English.

2. Approach

We assume that a key weakness of the neural approach implemented for TIAD-2019 was that it produces one single representation for a lexical entry, and that translations are identified by their distance from that representation. While this is a robust and reliable strategy for the most frequent sense of a particular word, we expect it to be less effective for polysemous or homonymous words, and to fail for rare and specialized senses. Indeed, identifying and classifying such senses, e.g., the use of a term in a particular domain of science, is a core tasks of lexicography, and part of the motivation for manual labor. We expect that KDictionary data is substantially richer in that regard that Apertium data.
Hence, instead of projecting embeddings for lexical entries, we project lexical concepts as identified in monolingual lexical resources. The most prominent, and widely used family of resources for this regard are word nets. Our approach on concept projection is based on WordNet synsets (Fellbaum, 1998). If source language and target language use the same synset identifiers (concepts), the target language translations of a particular source language word can be extrapolated from concepts by returning the most representative target language words for the concept associated with the source language. The challenge here is to develop metrics that express and maintain confidence of the association between a word and a concept. Based on these metrics, thresholds can be used to limit the set of possible concepts for a lexeme and possible lexicalizations of a concept. We employ the following core metrics:

- $P(\textit{concept}|\textit{lexeme})$: probability of a concept for a given lexeme (source or target word)
- $P(\textit{lexeme}|\textit{concept})$: lexicalization probability of a lexeme (source or target word) for a given concept
- $P(\textit{target}|\textit{source})$: translation probability of a target word for a particular source word
- $P(\textit{source}|\textit{target})$: translation probability of a source word for a particular target word

To initialize these metrics, we do not employ external resources to estimate them, but rather derive them from the branching factor within a WordNet, resp. a bilingual dictionary:

- *P(concept|lexeme)* := $\frac{1}{\textit{concepts(lexeme)}}$, where *concepts(lexeme)* is the number of concepts for a particular lexeme.
- *P(lexeme|concept)* := $\frac{1}{\textit{lexemes(concepts)}}$, where *lexemes(concept)* is the number of lexemes for a particular concept.
- *P(target|source)* := $\frac{1}{\textit{targets(source)}}$, where *targets(source)* is the number of target language words that the source word can be translated to (for a particular pair of source and target languages).
- *P(source|target)* := $\frac{1}{\textit{sources(targets)}}$, where *sources(targets)* is the number of source language words that the target word can be used for as translation (for a particular pair of source and target languages).

1. Initialization: For every (seed language) word that has a WordNet entry, assign its synset IDs as concepts as well as $P(\textit{concept}|\textit{lexeme})$ and $P(\textit{lexeme}|\textit{concept})$ scores.
2. First generation projection: For every *source* word without associated concepts that does have a translation relation to one or more *target* words (with associated concepts), calculate the *concept* probabilities as follows:

$$P(\textit{source}|\textit{concept}) := \sum_{\textit{target}} P(\textit{source}|\textit{target})\ P(\textit{target}|\textit{concept})$$

$$P(\textit{concept}|\textit{source}) := \sum_{\textit{target}} P(\textit{concept}|\textit{target})\ P(\textit{target}|\textit{source})$$

3. Iterate projection (second generation projection), until no more source words with translation relations to target words with associated concepts can be found.

In second and later generations, this procedure leads to a large number of low-probability associations between lexemes and concepts. To explore whether this has a negative effect, we also implemented a constrained variant (parameter `-constrained`): During projection, only those links between a lexeme and a concept are preserved that have maximum score (*s*=*source*, *c*=*concept*):

$$P(s|c) \mapsto \begin{cases} 0 & \exists k.P(s|c) < P(s|k) \\ P(s|c) & \text{otherwise} \end{cases}$$

$P(c|s)$ analogously

Using concept and lexicalization probability, translation inference (i.e., prediction *pred* of a target language word for a given source language word $source$) basically boils down to the following selection procedure:

$$\textit{pred} = \operatorname*{argmax}_{\textit{target}} \sum_{\textit{concept}} P(\textit{target}|\textit{concept})P(\textit{concept}|\textit{source})$$

We deviate from this trivial model as we aim to produce one prediction per concept, for a number of concepts with high values for $P(concept|source)$. In many cases, we found plain probabilities as extrapolated from the graph (we use no external resources except for concept inventories) to be indistinctive, so we coupled concept probability and lexicalization probability:

$$\textit{pred} = \operatorname*{argmax}_{\textit{target}} \sum_{\textit{concept}} P(\textit{target}|\textit{concept})P(\textit{concept}|\textit{target})\ P(\textit{concept}|\textit{source})\ P(\textit{source}|\textit{concept})$$

The intuition behind this term is that we return translation pairs that are optimal for every concept in both translation directions (from source to target language and vice versa). Algorithmically, we did not return the maximum value, but multiple translations, so we work directly with score metrics for a particular source word *source*:

$$score(concept) = P(concept|source)P(source|concept)$$

Accordingly, the score for a translation candidate $target$ is:

$$score(target) = \sum_{concept} P(target|concept)P(concept|target) \; P(concept|source) \; P(source|concept)$$

For translation prediction, we adopt the following selection procedure:

1. For translating the word *source*, retrieve the list of candidate concepts $C = \{concept|P(concept|source) > 0\}$ and the list of candidate translations $T = \{target|\exists concept \in C.P(target|concept) > 0\}$.
2. Sort C for decreasing $score(concept)$, sort T for decreasing $score(target)$
3. Optional: Restrict C to the first m elements (parameter `-maxConcepts`)
4. Optional: Enforce minimum concept score κ (parameter `-minConcScore`), i.e., eliminate all concepts from C with $score(concept) < \kappa$. The first element of C is maintained.
5. Initialize the result set R with the maximum lexicalization(s), i.e., lexicalizations with scores identical to that of the first element in T: $R = \{target|score(target) = score(t_1)\}$
6. For every element c_i in C, and those lexicalizations of c_i that are not R, add the lexicalization $target$ with maximum $P(target|concept)$ score to R. For candidates with identical $P(target|concept)$, return the target with maximum $score(target)$, i.e., maximum $P(concept|target)$, the highest degree of specificity.
7. Optional: Enforce minimum lexicalization score τ (parameter `-minLexScore`), i.e., eliminate all predictions $target$ from R with $score(target) < \min(\tau, score(t_1))$
8. Iterate in step 6 until no more lexicalizations are being added. Optional: limit iterations to n (parameter `-maxLexPerConcept`)

This procedure has a considerable number of parameters:

- The concept inventory being used
- unrestricted or constrained (`-constrained`) projection
- m (`-maxConcepts`): maximum number of concepts considered for translation inference
- n (`-maxLexPerConcept`): maximum number of lexicalizations per concept
- κ (`-minConcScore`): minimum $score(concept)$ for concepts considered during translation
- τ (`-minLexScore`): minimum $score(target)$ for possible translations

For every translation $target$, we return $score(target)$. For the official evaluation, the task organizers applied an additional threshold of 0.5 onto these values. As the aggregate diagram in Figure 1 does, however, show, our systems perform best (in terms of F1) without this additional threshold.

For our TIAD submission, this feature space was partially explored only, and it is likely that the KDictionary dictionaries used for evaluation require a different setting from the Apertium dictionaries that we take as input. As mentioned above, the Apertium dictionaries are designed for machine translation, so they are optimized for capturing the most frequent translation(s), whereas KDictionaries are designed for educational purposes, so they are optimized for capturing the most precise definition of words. In consequence, it is possible that a larger m and a lower κ score lead to better results on KDictionary data than they do on Apertium data. Our primary goal was thus not to fine-tune our systems to the Apertium data, but instead, to assess the contribution of concept inventories on translation inference across dictionaries.

3. Data & Preprocessing

We use the tab-separated value (TSV) edition of the dictionaries provided by the task organizers. Whereas we only use the languages and language pairs provided in these dictionaries, it would be possible to add more language pairs to be processed by our approach, as long as they are available in the TIAD-TSV format. We provide such data for more than 1,500 language pairs as part of the ACoLi dictionary graph (Chiarcos et al., 2020),[3] but this has not been considered in this experiment.

As for concept inventories, we use WordNet data, and we expect it to come as TSV data in accordance to the Open Multilingual WordNet specifications (Bond and Foster, 2013, OMW),[4] i.e., a three-column table containing synset ID in the first column, the string 'lemma' (or other relation identifiers) in the second column, and the word form in the third column. As for the word form, we differ from the OMW format by requiring that it is a Turtle string with a language tag, e.g., `"able"@en` instead if `able` in the English OMW WordNet. For OMW data, we provide a script that adds quotes and BCP47 language tags. We also provide a converter that produces OMW TSV from the RDF edition of Princeton WordNet 3.1.

A key advantage of OMW data is that it provides cross-linguistically uniform synset identifiers, so that multiple

[3] `https://github.com/acoli-repo/acoli-dicts`

[4] `http://compling.hss.ntu.edu.sg/omw/`

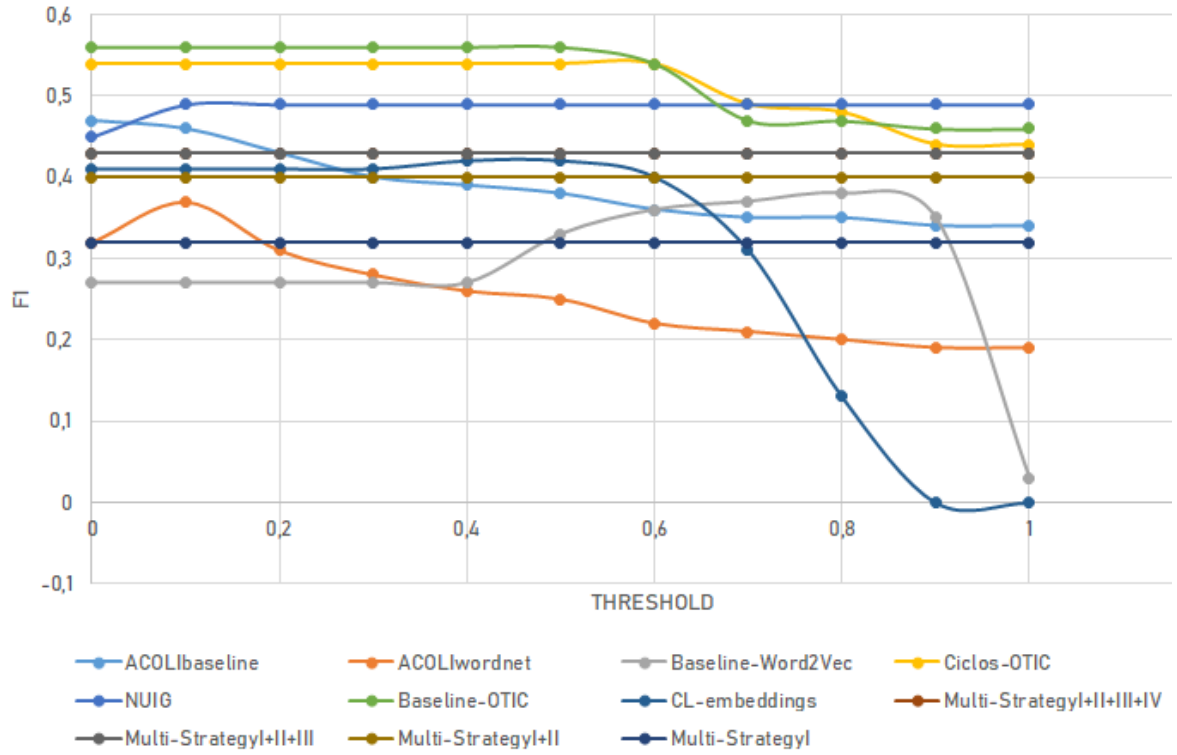

Figure 1: Official systems results (F1) per threshold.

WordNets can be combined for concept induction,[5] and our experiments included concept projection from multiple languages. The submitted system was based on the respective target language edition of OMW dictionaries (e.g., French for predicting English translations for French), as our internal experiments (see Tab. 3 below) indicated that combining multiple OMW dictionaries can produce worse results. Aside from projecting WordNet synsets, we also provide a baseline system that uses target language expressions instead of concepts (i.e., it is initialized with every target language expression being mapped to itself as a concept). The objective for doing so is to provide and to evaluate a knowledge-poor approach and also to evaluate the potential benefits that WordNet synsets might entail for this task.

4. Evaluation and Extensions

4.1. TIAD-2020 results

Based on the internal parameter optimization, we submitted results for the configurations summarized in Table 1, with the full Apertium graph as training data. Aiming for a typical number of translations per pair, we limited our predictions to the five highest-scoring translations. For achieving the reported results on precision (P), recall (R), F1 (F) and coverage (C), the task organizers applied an additional threshold of 0.5. Aside from the formal evaluation, they also provide the average results for variable thresholds, indicating that our systems perform better without an additional threshold (Table 2).

Both in our internal evaluation and in the official results, we found that using WordNet synsets for translation inference leads to a substantial decrease of translation quality in comparison to our baseline system that just projects translations. In terms of F1 measure, and without an additional threshold, this baseline system performs second among participant systems, whereas the WordNet-based system (in all configurations tested by the task organizers) ranks among the last three.

This may not be the last word on the usefulness of WordNet for translation inference across dictionaries, but it indicates that WordNet synsets are probably too coarse-grained for this task, so that relevant lexical distinctions are lost.[6] This may be compensated by corpus information about concept and lexicalization frequency, or, alternatively, by distributional methods to assess the prototypicality of a lexeme for a synset, e.g., the cosine similarity between word embeddings and synset embeddings as produced by (Rothe and Schütze, 2017). This approach can be a road to be explored in the future. For the moment, the intermediate summary is that projection-based translation inference performs better when translations are directly projected. It is conceivable to have better performance when word senses are projected, rather than synsets, but then, elaborate statistics about word sense frequencies would be necessary to select among projected word senses – that we do not possess at the moment.

[5] In principle, OMW synset identifiers can be used to generate translations *without any additional dictionary data*, for our TIAD-2020 contribution, we excluded the respective target language WordNet from projection experiments.

[6] As the high coverage of the ACoLi WordNet with threshold 0.0 (Tab. 2) shows, the drop in recall in comparison to the ACoLi baseline configuration is not the result of insufficient lexical coverage in the respective WordNets.

source	pt	pt	en	en	fr	fr
target	en	fr	fr	pt	en	pt
ACoLi WordNet, unconstrained, threshold 0.5						
WordNet	OMW Portuguese		OMW English		OMW French	
m	0.0	0.0	0.0	0.0	0.0	0.0
n	0.0	0.0	0.0	0.0	0.0	0.0
κ	∞	∞	∞	∞	∞	∞
τ	∞	∞	∞	∞	∞	∞
P	0.67	0.6	0.54	0.59	0.66	0.62
R	0.16	0.12	0.13	0.18	0.2	0.15
F	0.25	0.2	0.21	0.28	0.31	0.24
C	0.34	0.17	0.28	0.34	0.31	0.23
ACoLi Baseline, unconstrained, threshold 0.5						
m	0.0	0.0	0.0	0.0	0.0	0.0
n	0.0	0.0	0.0	0.0	0.0	0.0
κ	∞	∞	∞	∞	∞	∞
τ	∞	∞	∞	∞	∞	∞
P	0.66	0.64	0.48	0.57	0.63	0.63
R	0.26	0.26	0.24	0.3	0.32	0.27
F	0.38	0.37	0.32	0.39	0.42	0.38
C	0.54	0.36	0.51	0.55	0.48	0.42
best-performing participant system per language						
	Ciclos-OTIC	ACoLi Baseline	Ciclos-OTIC	NUIG	ACoLi Baseline	Ciclos-OTIC
F	0.53	0.37	0.5	0.49	0.42	0.6
C	0.79	0.36	0.74	0.55	0.48	0.74
baseline	OTIC	OTIC	OTIC	OTIC	OTIC	OTIC
F	0.51	0.72	0.48	0.53	0.48	0.62
C	0.76	0.8	0.68	0.71	0.54	0.72

Table 1: Official TIAD-2020 results per language

system	P	R	F	C
better-performing participant systems (wrt. F1 score)				
Ciclos-OTIC	0.64	0.47	0.54	0.76
NUIG	0.77	0.35	0.49	0.54
ACoLi baseline, unconstrained (best threshold and official threshold)				
0.0	0.37	0.64	0.47	0.96
0.5	0.60	0.28	0.38	0.48
ACoLi WordNet, unconstrained (best thresholds and official threshold)				
0.0	0.22	0.64	0.32	0.96
0.1	0.52	0.28	0.37	0.48
0.5	0.61	0.16	0.25	0.28
baselines (with thresholds)				
W2V (0.8)	0.48	0.32	0.38	0.59
W2V (0.5)	0.30	0.37	0.33	0.68
OTIC (0.5)	0.69	0.48	0.56	0.71

Table 2: TIAD-2020 evaluation results: Averaged scores for systems with variable threshold

4.2. Constrained concept projection

For the official TIAD evaluation, we submitted systems with unconstrained concept projection only. In a follow-up experiment, we also evaluated constrained projection as a promising direction to counter the weakness of WordNet-based concept projection. This may indeed be the case, as we could confirm that constrained projection systematically outperforms unconstrained projection of WordNet synsets. For this evaluation, we replicated the TIAD evaluation setting by aiming to predict an Esperanto-English

dictionary (and excluding it from the training data). Our evaluation setup differs from TIAD evaluation in that we do not exclude out-of-vocabulary words (that cannot be predicted from other dictionaries). Table 3 summarizes the overall results, but even in this configuration, direct projection of translation outperformed concept-based translation inference. As a side-observation, we found that using multilingual WordNets can have a negative impact on precision, possibly because of imprecisions in the alignment between multilingual synsets. We also found that concept projection from *selected* dictionaries (here Esperanto-Spanish and Spanish-English only) may lead to slightly better F1 scores than projection over the *full* set of dictionaries. However, it is not clear whether this represents a factual improvement, as it is naturally accompanied with a lower degree of coverage (not reported in the table).

4.3. Comparison with embedding projection

Our second objective was to evaluate concept-based translation inference in comparison with the embedding projection we provided to TIAD-2019. Unfortunately, the results are not directly comparable, so that they can be evaluated only for the internal evaluation setup also applied for constrained concept projection.

Our TIAD-2019 system employed a simple technique for projecting word embeddings from a seed language (or, if multilingual embeddings are available, from multiple seed languages) over the translation graph. For a given word in the source language, with embedding $\vec{s}$, its translations are predicted from the cosine similarity between $\vec{s}$ and the vectors of translation candidates $\vec{t_{1..n}}$ in the target languages. Aside from seed languages and the original word embeddings, its main parameters are the number of translation candidates returned (`-maxMatches`) and a minimum similarity threshold applied to the predictions (`-minScore`). We used English as a seed language, with the same vectors (50-dimensional GloVe embeddings) as in our TIAD-2019 submission.

The results are summarized in Tab. 4: We report the best-performing configurations for `-minSimilarity=0.0` (`-maxMatches` $\in \{1, ..., 5\}$), `-maxMatches=5` (`-minSimilarity` $\in \{0.0, 0.1, ..., 1.0\}$) and `-minSimilarity=0.9` (`-maxMatches` $\in \{1, ..., 5\}$). One important observation here is that the best-performing configuration is one that limits the number of predicted translations to 1, indicating that the neural model performs best for predicting the translations based on the *main* sense. In other words, the Apertium dictionaries seem to avoid additional synonyms for synonymous target language translations of a given source word, but to provide alternative translations to express target language translations that relate to different source language senses (and are not synonymous in the target language).

The question is now whether WordNet concepts can be used in a meaningful manner to provide translations for secondary senses. In line with the findings of the TIAD-2020 evaluation, the unconstrained WordNet systems basically fail and are outperformed by direct translation projection ('ACoLi baseline') by a large margin. However, this is not the case for *constrained* WordNet systems that reach (or, depending on configuration, beat) the neural baseline. This indicates that our approach is indeed capable of preserving lexicographically relevant sense distinctions. The overall best-performing system is, however, not based on WordNet concepts, but on direct translation projection.

Furthermore, we find that projection is an effective approach only if it is limited. Constrained projection generally produces better results, in particular for WordNet concepts, and the additional filters that κ, τ, n and m provide can be employed to reach further, substantial, improvements over the vanilla systems we submitted to TIAD-2020. Although we cannot evaluate on TIAD task data directly, we see our approach as a promising direction for future participation in future tasks. In particular, we substantially outperform the best-performing TIAD-2020 system, the OTIC baseline provided by the task organizers.[7]

5. Discussion & Conclusion

In this paper, we described the vanilla implementation we provided for the Third Shared Task on Translation Inference Across Dictionaries, as well as a number of subsequently developed improvements to this system.

We developed our system in an attempt to address a likely source of shortcomings of our earlier TIAD-2019 system. We did not resubmit our TIAD-2019 system, however, because we expected the evaluation data to be identical. This is not the case, and the data may have different characteristics than the 2019 data, as the substantial boost in performance of the organizer baseline system systems indicate. Instead, we performed a comparative evaluation for our 2019 and 2020 systems on the EO-EN Apertium dictionary.

We assume that our 2019 system, based on the projection of embeddings for lexical entries over the translation graph, performs relatively well on capturing the most frequent sense, but that it fails for translation relations of secondary senses. We thus explored the possibility of projecting WordNet synsets over the translation graph, and using these for translation inference. In order to evaluate the effectiveness of synsets for this purpose, we also performed a baseline experiment where we projected translations instead of concepts. To our surprise, This baseline outperformed WordNet-based translation inference in all configurations.

This is also confirmed by the TIAD evaluation, albeit our baseline fares relatively well among the first three systems (with variable threshold) – the WordNet system does not.

In the internal evaluation, we also compared our 2019 system. In its vanilla configuration (with unrestricted projec-

[7] As for the comparably poor performance of the OTIC baseline in our setting in comparison to the TIAD-2020 blind evaluation, this seems to be due to a coverage issue. We ran the evaluation over the entire Esperanto vocabulary in the Apertium graph. However, when out-of-vocabulary words are excluded from the evaluation, i.e., words for which no pivot language translation can be found, OTIC (pivot Spanish, threshold 0.5) yields precision 0.67, recall 0.62, and F1 0.65, roughly corresponding to the TIAD-2020 scores of the OTIC system. Another difference in our evaluation was that we did not distinguish homonyms with different part of speech tags.

constrained	κ	τ	n	m	WordNet	dictionaries	P	R	F
no	0	0	∞	∞	none	all	0.26	0.32	0.29
yes	0	0	∞	∞	none	all	0.52	0.25	0.33
no	0	0	∞	∞	none	EO-ES-EN	0.63	0.22	0.33
yes	0	0	∞	∞	none	EO-ES-EN	0.68	0.18	0.28
no	0	0	∞	∞	en	EO-ES-EN	0.10	0.24	0.14
yes	0	0	∞	∞	en	EO-ES-EN	0.48	0.19	0.27
no	0	0	∞	∞	en	all	0.03	0.36	0.05
yes	0	0	∞	∞	en	all	0.34	0.27	0.30
no	0	0	∞	∞	all*	all	0.06	0.43	0.11
yes	0	0	∞	∞	all	all	0.22	0.33	0.26

*all WordNets: ca, en, es, eu, gl, it, pt

Table 3: Evaluating constrained projection

system configuration					evaluation		
ACoLi-neural, GLoVe 6B (TIAD-2019 system)							
-maxMatches	*-minScore*	*seed language*	*embeddings*	*length*	*P*	*R*	*F*
1	0.0	en	GloVe 6B	50	0.67	0.22	0.33
5	0.9	en	GloVe 6B	50	0.58	0.22	0.32
2	0.9	en	GloVe 6B	50	0.63	0.22	0.33
translation projection ('ACoLi baseline', unconstrained)							
m	*n*	κ	τ	*WordNet*	*P*	*R*	*F*
∞	∞	0.0	0	none	0.26	0.32	0.29
3	1	0.3	0	none	0.55	0.47	**0.51**
translation projection ('ACoLi baseline', constrained)							
∞	∞	0.0	0	none	0.52	0.25	0.34
3	1	0.3	0	none	0.58	0.43	**0.49**
concept projection ('ACoLi WordNet', constrained)							
∞	∞	0	0	en	0.34	0.27	0.30
3	2	0.3	0	en	0.44	0.44	0.44
ACoLi TIAD-2020 systems (unconstrained)							
ACoLi Baseline				none	0.26	0.32	0.29
ACoLi WordNet				en	0.03	0.36	0.05
OTIC (TIAD-2020 best-performing system, default and best threshold)							
configuration			*pivot language*	*-minScore*	*P*	*R*	*F*
default threshold			Catalan	0.5	0.65	0.19	0.30
best threshold			Catalan	0.2	0.59	0.21	0.31
default configuration			Spanish	0.5	0.67	0.21	0.32
best configuration			Spanish	0.0	0.64	0.23	0.33

Table 4: Comparing TIAD-2020 baseline, concept projection, translation projection and embedding projection techniques for predicting the Apertium EO-EN dictionary (best-performing configurations).

tion), the ACoLi baseline also falls behind that. However, in an extension of our TIAD system that implements *constrained projection*, where only the highest-scoring lexicalization and concept probabilities are preserved, lead to better F1 scores, and further improvements can be achieved if concept (translation) projection is limited to a low number of translation candidates (3) and further confidence thresholds are applied. The improvements bring both concept projection and translation projection approaches to the performance of original embedding projection technique, and the overall best-performing system (in our internal evaluation) is a configuration of the translation projection approach.

Our system is both simple and knowledge-poor. It does not require any multilingual data beyond bilingual dictionaries, and it can be applied (apparently with even better results) without monolingual sense inventories. Obviously, this is a natural starting point for further extensions. We extrapolate translation probabilities, concept probabilities and lexicalization probabilities only from the structure of the lexi-

cal resource(s), but empirical frequency measurements and other corpus-derived information may provide a much more accurate picture (for the effective use of a lexeme, at least, although maybe less so for its lexicographic characteristics). In particular, future directions may include a combination of neural and concept-based approaches. As such, translation inference from projected synsets may be more robust and the coverage may improve if lexicalization is not directly based on WordNet, but if the distributional similarity between target language words and synsets is used as a measurement of lexical prototypicality of a word for a concept. Such an approach requires synset embeddings that reside in the same feature space as the corresponding word embeddings, and indeed, this would be possible with techniques for inducing synset embeddings, e.g., as described by Rothe and Schütze (2017).

Another possible extension is to combine our approach with the OTIC baseline. In our internal evaluation, the OTIC baseline suffered from coverage issues in the pivot language dictionaries. It was thus outperformed by the projection-based approach as this takes the entire source and target language vocabulary provided by the Apertium dictionary graph into consideration. Future experiments may adopt OTIC for source language lexemes that do have a pivot language translations and use concept, translation or embedding projection for out-of-vocabulay elements.

6. Acknowledgements

The research of the first and third author of this paper on this topic was financially supported by the project "Linked Open Dictionaries" (LiODi, 2015-2020), funded by the German Ministry for Education and Research (BMBF) as an Independent Research Group on eHumanities. The research of the second and third author was financially supported by the Research and Innovation Action "Prèt-a-LLOD. Ready-to-use Multilingual Linked Language Data for Knowledge Services across Sectors" funded in the European Union's Horizon 2020 research and innovation programme under grant agreement No 825182.

We would like to thank the anonymous reviewers for helpful comments and feedback, and the TIAD-2020 shared task organizers and Marta Lanau Coronas for providing the code of the TIAD-2020 OTIC baseline.

7. References

Bond, F. and Foster, R. (2013). Linking and extending an open multilingual wordnet. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL-2013)*, pages 1352–1362.

Chiarcos, C., Fäth, C., and Abromeit, F. (2020). Annotation interoperability in the post-ISOCat era. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020)*, Marseille, France. Accepted for publication.

Donandt, K. and Chiarcos, C. (2019). Translation inference through multi-lingual word embedding similarity. In J. Gracia et al., editor, *Proc. of TIAD-2019 Shared Task Translation Inference Across Dictionaries (`http://ceur-ws.org/Vol-2493/`) at 2nd Language Data and Knowledge (LDK) conference.*

Christiane Fellbaum, editor. (1998). *WordNet: An Electronic Lexical Database*. Language, Speech, and Communication. MIT Press, Cambridge, MA.

Pennington, J., Socher, R., and Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, October. Association for Computational Linguistics.

Rothe, S. and Schütze, H. (2017). AutoEextend: Combining word embeddings with semantic resources. *Computational Linguistics*, 43(3):593–617.