The GuanXi network: a new multilingual LLOD for Language Learning applications

Ismail El Maarouf[1] **, Eugene Alferov**[2]**, Doug Cooper**[3]**,**
Zhijia Fang[4]**, Hatem Mousselly-Sergieh**[5]**, Haofen Wang**[6]
[1]University of Wolverhampton, United Kingdom. [2]Kherson State University, Ukraine.
[3]CRCL, USA. [4,6]East China University of Science and Technology, China.
[5]Ubiquitous Knowledge Processing Lab, Technische Universität Darmstadt, Germany.
[1]i.el-maarouf@wlv.ac.uk, [2]alferov.evgeniy@gmail.com,
[3]doug.cooper.thailand@gmail.com,[4]fantorm030@gmail.com,
[5]mousselly-sergieh@ukp.informatik.tu-darmstadt.de,
[6]whfcarter@ecust.edu.cn

Abstract

Linguistic resources are essential for Language Learning applications. However, available resources are usually created in isolation, thus, they are scattered and need to be linked before they can be used for a specific task such as learning of a foreign language. To address these problems we present a new resource that link linguistic resources of multiple languages using the framework of Linguistic Linked Open Data (LLOD).

1 Introduction

This paper presents the GuanXi[1] network, a multilingual Linguistic Linked Open Data (LLOD) resource. GuanXi is to be integrated in a language learning platform to provide course designers with easy access to quality language data on a variety of media (text, audio, video, image) in order to support the construction of learning activities, but also harvest the power of Linked Data to suggest new views on data, as well as new activities.

For this particularly sensitive application, the GuanXi network provides reliable linked data where links are of high quality. GuanXi currently focuses on verbs and draws on recent RDF conversions of various LLOD such as PDEV-lemon (El Maarouf et al., 2014), Slovnyk and COW (Wang and Bond, 2013).

This paper presents this network and the methods used to build it and evaluate the multilingual sense links. The work presented here focuses on techniques where WordNet[2] is used as an interlingual index, and where corpus data can be leveraged, integrated, and connected to the lexical entries at the level of sense. Corpus data is particularly important for language learning as it provides massive amounts of real language use.

Section 2 describes related work on resources and technologies of sense linking. Section 3 presents the resources integrated in the GuanXi Network. Section 4 presents the different methods used to build the GuanXi network for each language pair, depending on available resources and section 5 presents the data model using the LLOD framework. Section 6 provides both automatic and manual evaluations of sense linking strategies and section 7 concludes on future work.

2 Related Work: sense linking

A major concern of Linked Data (LD) is to meaningfully interconnect resources in a way that is consistent and reliable. For Linguistic LD (LLD), this implies that introducing links at the level of the sense is of a much higher quality and usefulness than at the level of, say, the entry. This is because each lexical entry may offer a number of senses and, since words can be polysemous, getting the sense wrong will lead to disastrous consequences or limited progress, for any application that makes use of the resource. It is important to note that this is not specifically an issue of LLD, but of language processing in general and semantics. Overall, linking entities belonging to two different resources consist in automatically extracting existing information relevant to each entity within each resource and compute a similarity for each possible link.

Methods include aligning senses of different resources (e.g. WordNet and FrameNet) based on the similarity of the corresponding glosses/definitions. This technique was used in UBY (Gurevych et al., 2012; Niemann and Gurevych, 2011) where the alignment between

[1]Literaly, guanxi, or 关系, is Chinese for relationship.
[2]http://wordnet.princeton.edu/

Proceedings of the Second Workshop on Natural Language Processing and Linked Open Data, pages 42–51,
Hissar, Bulgaria, 11 September 2015.

two senses is determined based on the cosine similarity of their gloss representations. Another family of approaches for word-sense alignment uses graph methods, such as personalized page rank (PPR) (Agirre and Soroa, 2009), Dijkstra-WSA (Matuschek and Gurevych, 2013) and BabelNet (Navigli and Ponzetto, 2012).

Techniques for aligning senses from resources of different languages have also been proposed, mainly by applying Machine Translation to get translated glosses, and compute in a second step the similarity. This is, for instance, the method used in UBY to connect OmegaWiki and WordNet (Gurevych et al., 2012; Bond and Foster, 2013). Because these methods rely on definitions, they are very similar to Lesk similarity variants in Word Sense Disambiguation (WSD) (Lesk, 1986; Banerjee and Pedersen, 2002), which compute the similarity between a definition and an example in order to assign the correct sense.

Following that, methods making use of corpus data have been proposed. BabelNet is the result of (among other things) harvesting sense-tagged corpora and their automatic translation by Google Translate of WordNet annotated SemCor and Wikipedia (Navigli and Ponzetto, 2012). Babelnet also makes use of graph-based methods (Mihalcea, 2005; Navigli and Ponzetto, 2012).

BabelNet contains lexical data for over 270 languages and can be accessed through a WSD service, named Babelfy, which automatically annotates the sense of each content word in a sentence from any of the 270 languages. Babelfy uses a unified graph-based approach that combines Event Linking (EL) and WSD techniques. Given a text that should be disambiguated, all linkable fragments are extracted and for each fragment, a list of a candidate senses is extracted according to a semantic network. The semantic network contains a signature for each concept, that is, a set of related concepts. Next,a graph-based semantic interpretation for the input text is created, by linking the candidate senses of the extracted fragments using the previously-computed semantic signatures. Finally, a dense subgraph of this representation is extracted and the best candidate sense for each fragment is selected.

However, the techniques described in this section have unsatisfying accuracy, as much of the information is missing, and (automatic) Word Sense Disambiguation is still not solved (Kilgarriff and Palmer, 2000; Navigli, 2009), and is generally around 70% accuracy. The best way to link linguistic data accurately therefore still depends ultimately on lexicographical expertise. This is, for instance, the approach taken in WordNets (Bond and Paik, 2012).

Using lexicographic expertise to identify sense links should avoid (resource) publication bias, experiments and resources bootstrapping on the same data over and over again, and will open new perspectives. Note that using lexicographic expertise does not mean that automatic methods should be discarded; in fact the approach described in this paper makes use of semi-automatic methods for dataset linking, and lexicographer input is kept to the evaluation stage of the cycle. This paper explores the idea that the main concern for accurate LLD is to design efficient frameworks to make the best use of Human expertise in a minimum of time.

3 Target Resources

In aligning lexical resources, WordNet is almost inescapable as the English WordNet is manually connected to several languages (but see (Sérasset, 2012), for a different approach). However, comparatively few resources/languages are connected to WordNet. Even BabelNet has limited coverage for languages which are less resourced than English (e.g. Ukrainian). Moreover, other lexical resources exist even for English that contain valuable knowledge but are not connected. This section provides a short description of the resources used in this paper.

3.1 The Pattern Dictionary of English Verbs (PDEV)

PDEV [3] is a dictionary of English verbs. It is based on a new technique, called Corpus Pattern Analysis (CPA)(Hanks and Pustejovsky, 2005; Hanks, 2012; Hanks, 2013; Baisa et al., 2015), for mapping meaning onto words in text. CPA is also influenced by frame semantics (Fillmore, 1985) and PDEV can be seen as complementary to FrameNet[4]. Where FrameNet offers an in-depth analysis of semantic frames, CPA offers a systematic analysis of the patterns of meaning and use of each verb. Each CPA pattern can in principle be plugged into a FN semantic frame. In PDEV verb patterns consist not only of the basic "argument

[3] http://pdev.org.uk

[4] https://framenet.icsi.berkeley.edu/

structure" or "valency structure" of each verb, but also of subvalency features, where relevant, such as the presence or absence of a determiner in noun phrases constituting a direct object. Each argument in a PDEV pattern is populated with Semantic Types (taken from a shallow semantic ontology[5]) indicating the preferred semantic set of entities which are prototypically found in each slot. PDEV is a unique resource in this regard. It is also the output of a corpus-based lexicographical approach and provides extensive sets of examples from real language data.

PDEV has recently been converted into RDF (El Maarouf et al., 2014) using the lemon model (McCrae et al., 2011). PDEV-lemon contains 17,634 triples, 3,702 patterns/senses for 984 entries and the dump obtained for this paper covers an up-to-date lexicon of 1,273 entries and 4,531 patterns/senses.

3.2 Chinese Open Wordnet (COW)

The Chinese Open Wordnet (COW) is a large scale, free dictionary for Mandarin Chinese (Wang and Bond, 2013). COW was created to address the main limitations of other Chinese WordNets, namely the coverage and the quality of the data. To achieve this, a three-phase procedure was applied:

1. data was extracted from the Wikitionary[6] and merged with SEW (Southwest University WordNet) (Xu et al., 2008),
2. manual check was performed on the translations, and
3. the semantic relations were also checked manually.

Currently, COW includes 42,315 synsets with 79,812 senses and 61,536 unique words.

3.3 Slovnyk Dictionary

Slovnyk[7] is a multilingual dictionary that supports bilingual translation among 32 languages. For a word in a source language, Slovnyk provides the corresponding translation in the target language according to the most common sense of the source word. In contrast to WordNet, Slovnyk does not provide grammatical information, sense information, or semantic relation between terms. In this paper, we obtained a subset of Slovnyk for two language pairs: English - Ukrainian, and Ukrainian - Spanish. This has been converted into RDF, with a separate lexicon for each language using the lemon model (McCrae et al., 2011), and a translation set for each language pair[8].

3.4 Apertium

As Slovnyk mainly contains nouns and noun phrases, we automatically extracted verbs from the Apertium Russian-Ukrainian bilingual lexicon[9]. Apertium (Corbí Bellot et al., 2005) was an open-source rule-based Machine Translation platform, which therefore heavily relies on bilingual lexica and grammars. It is now supported by an online community[10]. This method enables to collect 1,215 different verbs, which were integrated into the Slovnyk Ukrainian dictionary.

3.5 Corpora

We use two corpora in our experiments. The first is the British National Corpus (BNC) (Burnard, 2007), a large reference corpus of British English (100 million words). We use the version that is available through PDEV and because it is annotated with pattern numbers.

The second corpus is OPUS, an aligned multilingual corpus containing various sources for 92 languages (Tiedemann, 2009). We focus on the Ukrainian-English pair which contains movie subtitles and technical software documentation (3.3 million words) made available through the SketchEngine query system (Kilgarriff et al., 2014).

4 WordNet senses as interlingual links

We present a cross lingual approach to establish links between lexical semantic resources (LSRs) and corpora.

Our approach is fairly standard in this respect as it aims to use WordNet (WN) as a multilingual index between languages. This approach requires two steps:

1. identify appropriate WN senses for each sense in each resource
2. link all entry pairs with a sense in common

The resulting translation pairs are the pairs which have a WN sense in common. This method can be applied to Open Multilingual WordNet. In this experiment we use COW, the Chinese Open

[5] http://pdev.org.uk/#onto
[6] https://www.wiktionary.org/
[7] http://www.slovnyk.org/
[8] http://datahub.io/dataset/rdf-uk-es
[9] http://wiki.apertium.org/wiki/Russian_and_Ukrainian
[10] apertium.org

WordNet, which provides links between Chinese words and WN senses (manually checked). The general workflow is illustrated in Figure 1

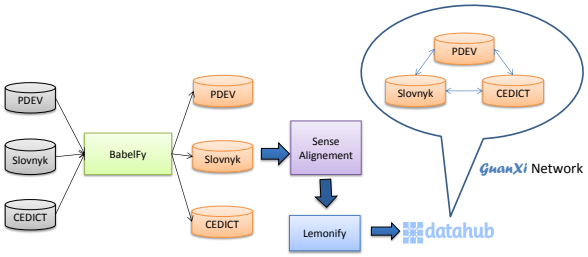

Figure 1: Approach workflow

4.1 Word Sense Disambiguation for Wordnet sense harvesting

PDEV is an isolated resource in the linked data cloud, so the links to WordNet need to be created. However it possesses its own sense-tagged corpus, which means that if the WordNet sense of the verb in one of these examples is correctly disambiguated, the pattern sense can be mapped to a WN sense. In order to do that, we mainly used the Babelfy API[11], which provides a disambiguation service that outputs a BabelNet sense for each content word. Since BabelNet builds on WordNet for verbs, the WN sense can straightforwardly be derived from the BabelNet sense. Thus all that is needed is an example of a sense from PDEV, in order for Babelfy to identify the relevant sense. This can be performed for any of the 271 languages covered by Babelfy.

However, this technique has its limits, since, as we discovered in our initial experiments, it is not possible to simply query Babelfy on any language and build bilingual lexica by collecting common senses. In fact, the languages targeted in the GuanXi network (Chinese and Ukrainian), have poor support and either need query pre-processing, or more lexical coverage. This is the reason why we made use of various data sources in combination: in order to link English with Chinese, we rely on the WN sense links provided in COW.

4.2 Beyond WordNet: Example-based sense mapping

For less resourced languages such as Ukrainian, which are not linked to WordNet, we propose an alternative example-based method, which consists in taking a non-English example and annotating relevant e.g. Ukrainian tokens with a pattern sense. Thus, we can harvest a sense for a Ukrainian verb (the pattern) and a link between English and Ukrainian (the translation) at the sense level.

The reason for using this approach is that we consider that PDEV patterns are very reliable representations of sentence meaning: as opposed to a standard definition or gloss, it specifies the contextual conditions of use in great detail, which is of great help to the annotator. Obviously, this will provide an incomplete picture of the language (since some senses which may be specific to a non-English language, with the consequence that finer-grained semantic preferences, may not be discovered with this technique), and cannot be used to identify translations which map to different parts of speech.

In this context, we can leverage examples from parallel corpora, which already provide translation candidates in context. This greatly decreases the workload on human annotation and provides a controlled framework for verb translation.

5 The Multilingual Corpus-Lexicon Model

5.1 Resources Types for Language Learning

The main type of resources that are connected in the GuanXi network are corpora, lexica and ontolgies (including taxonomies). Thanks to the concept of Linked Data, this network allows the extraction of multiple datasets resulting from different views on the network. Thus, it is possible to extract examples for senses, but also examples where a given semantic type is the subject of a verb, etc. Currently, only the verb token in the corpus example can be directly linked to lexical entries, but we intend to multiply annotations on examples in a semi-automatic way in order to enable the retrieval of other entities in each examples. Particularly we plan to include the Semeval 2015 dataset for Task 15[12], which includes annotations for 4,529 sentences, which can be straightforwardly mapped to both syntactic (syntactic relations) and semantic (semantic types) classes of PDEV-lemon.

We are particularly keen on using corpus examples because the end users of this resource, lan-

[11] http://www.babelfy.org/

[12] alt.qcri.org/semeval2015/task15

guage learners, need to work on/with real language use. PDEV provides the list of patterns that are most commonly used in English, i.e. those which a foreign speaker should learn in priority. In fact, it is possible to design a progressive learning curriculum, since PDEV provides percentages of uses of each pattern of each verb. PDEV also classifies examples according to whether they are normal pattern uses or creative and figurative uses. Selecting appropriate examples is therefore greatly facilitated by this prior massive manual work.

5.2 The GuanXi Framework

These resources can all be integrated into a data model. We use the lemon framework (McCrae et al., 2011) to represent the lexicons and the NIF model (Hellmann et al., 2013) to represent corpus data. Lemon has a relation for creating links between senses and examples but the example class is not structured. The ability to isolate a word from a sentence in order to refer to it, or to appropriately annotate a sentence part with links to features of an entry is instead provided by the NIF model. The main principle of lemon is to provide a model which enables the separation of lexical information from semantic information as provided in ontologies. The GuanXi network is connected to 8 ontologies and lexinfo[13]. Finally we use the translation[14] module described in (Gracia et al., 2014) as the translation framework for bilingual lexicons.

The resulting multilingual corpus-lexicon-ontology data model of the GuanXi is illustrated in Figure 2. As can be seen, we use a new relation *kwic* (Key Word In Context) to relate a particular token of a sentence in NIF representation with a lexical sense in a specific language. This link makes it possible to have a simple but powerful link between the corpus and the lexicon, without having to rely on external ontologies, in line with lemon principles. The translation set helps to connect various equivalent senses of words from different languages. The figure also shows the structure of PDEV verb entries and the links between the lexicon and the ontologies. It is worth noting that we only use the ontology part of FrameNet (the frame and frame elements), which is connected to a concept in the PDEV ontology.

[13] http://lexinfo.net/

[14] http://purl.org/net/translation.owl

6 Evaluation

6.1 Automatic evaluation through clustering similarity

The Babelfy system provides state of the art performance on Word Sense Disambiguation (WSD) (Moro et al., 2014). However, WSD systems can experience a significant drop in performance when evaluated on unseen data, and generally have very different results on different datasets.

Since the quality of the links of the GuanXi network depends on Babelfy's ability to identify the right BabelNet synset in context, we set up an experiment to automatically assess the quality of this disambiguation. Since each PDEV pattern is connected with a set of examples, we submitted these examples (to the maximum of 5 per pattern) for disambiguation to Babelfy and extracted the BabelNet synset.

In order to evaluate the quality of the mappings, we used the B-cubed definition of Precision and Recall, first used for coreference (Bagga and Baldwin, 1999) and later extended to cluster evaluation (Amigó et al., 2009). Both measures are averages of the precision and recall over all instances. To calculate the precision of each instance we count all correct pairs associated with this instance and divide by the number of actual pairs in the candidate cluster that the instance belongs to. Recall is computed by interchanging Gold and Candidate clusterings[15] (Eq. 1).

$$\text{Precision}_\text{i} = \frac{\text{Pairs}_\text{i} \text{ in Candidate found in Gold}}{\text{Pairs}_\text{i} \text{ in Candidate}}$$
$$\text{Recall}_\text{i} = \frac{\text{Pairs}_\text{i} \text{ in Gold found in Candidate}}{\text{Pairs}_\text{i} \text{ in Gold}} \quad (1)$$

Table 1 compares Babelfy with standards WSD algorithms such as Simple Lesk (Lesk, 1986) or Adapted Lesk (Banerjee and Pedersen, 2002)[16], taking into account the full sentence. Every system beats the baseline, Baseline1, which consists in assigning all examples the same sense (i.e. without account of context). According to B-cubed F-score, Adapted Lesk provides clusterings which are the most similar to PDEV.

[15] A clustering is the set of clusters that a particular method outputs.

[16] This study uses the pywsd implementation; for more details, see (Tan, 2014)https://github.com/alvations/pywsd

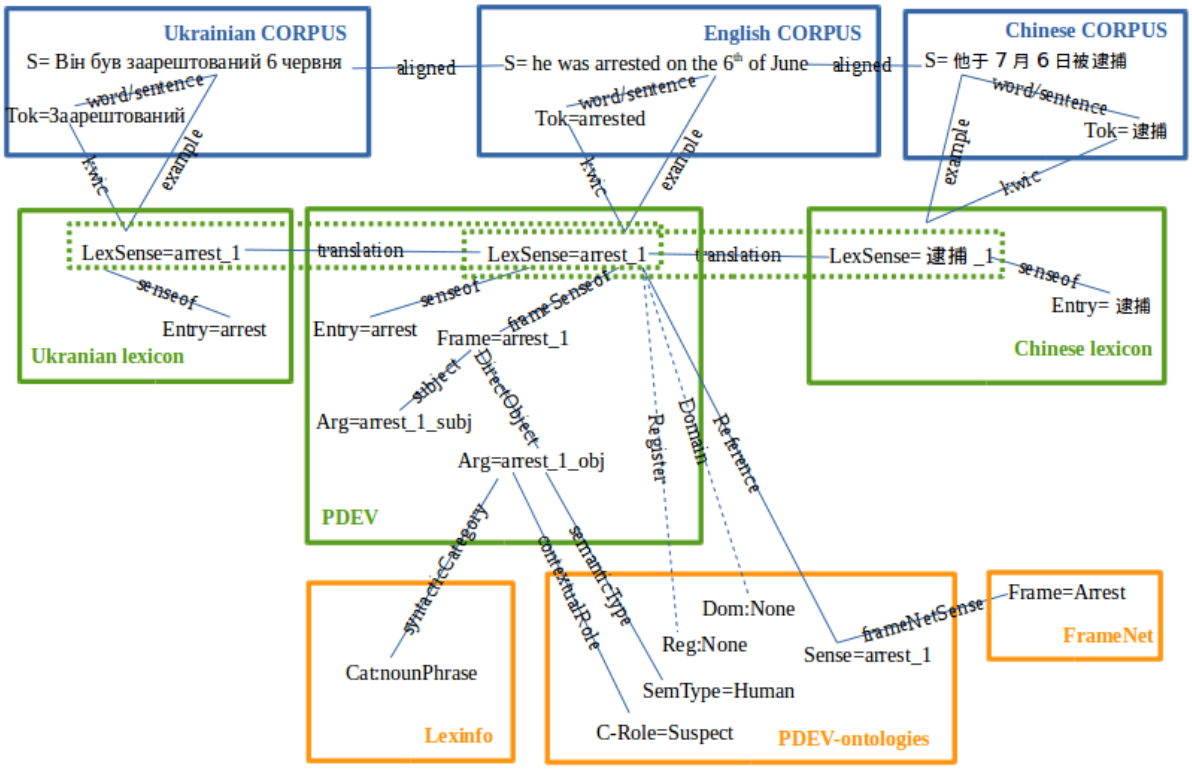

Figure 2: Guanxi Data Model

System	B^3 F-score
Cosine Lesk	0.470
Baseline1	0.472
Orig Lesk	0.579
Babelfy	0.639
Simple Lesk	0.655
Adapted Lesk	0.656

Table 1: Results for WSD on full sentence using B-cubed F-score.

Context	B^3 F-score
Size=1	0.633
Size=2	0.666
Size=3	0.668
Size=4	0.666
Size=5	0.662
Baseline1	0.472

Table 2: Optimising context size for Babelfy WSD using B-cubed F-score.

This evaluation calls for two warnings. First, evaluating the clusterings of two methods or resources tells in theory nothing about the quality of these clusters: a system might well cluster tokens identically to the reference but provide wrong pointers to WordNet definitions or senses. However in practice, assuming that good clusterings gives a strong indication of quality is a reasonable assumption.

The second warning is that clusters obtained from PDEV do not necessarily signal sense differences. Therefore the algorithm might well be correct in assigning to 2 different patterns, one and unique WordNet sense. However, as opposed to other clustering evaluation measures (see e.g. Measure Of Concordance (Pfitzner et al., 2009)), the B-cubed measure tends to attenuate the impact of this kind of cases.

We decided to use the Babelfy system for WSD, mainly because it returns BabelNet synsets, thereby providing access to many resources, and because previous evaluations have shown the effectiveness of the algorithm. So we proceeded to optimize the query system by identifying the best size for a query.

We submitted six sets of queries: queries which included one context word (on each side of the target word) in addition to the target word, but also two, three, four, and five context words, and the full example. Table 2 shows that the optimal size

of context for Babelfy is 3 words on each side of the target word, and no benefit is obtained by taking into account more context; on the contrary, the performance tends to decrease, to the point that it is almost equivalent to a context size of 1.

6.2 Chinese manual evaluation

We generated a small corpus of examples for each PDEV pattern, with maximum 5 examples per pattern. This covered 4,532 patterns of 1,274 verbs. We used Babelfy with the best setup (+-/3 words) to get the WN synsets.

Out of 19,651 English queries, Babelfy returned links to WordNet except for 279 examples (NA), and 216 "null" WN senses (95 verbs, 88 nouns, 30 adj, and 3 adverbs), meaning a coverage rate of 97.5% (4,469 patterns for 1,240 verbs). With respect to null verb synsets, these are senses from the Wiktionary that have not been mapped to a WordNet synset. For example, *rewind* with the gloss *to wind (something) again*, also exists in WordNet with the gloss *rewind (wind (up) again) 'the mechanical watch needs rewinding every day'*. Example (1) illustrates a case of NA concerning verb *abduct*, which is probably due to a processing error or threshold on the Babelfy API.

(1) Police believe he died a few hours after he was abducted .

Our version of COW contains 80,010 word-synset pairs, covering 61,535 Chinese words and 42,315 English synsets. Out of these, 1,214 COW different links to WN overlapped with those obtained with Babelfy. 10,796 examples (55%) could be matched with a common WN synset, covering 2,918 patterns (65%) for 807 entries (65%).

We then proceeded to evaluate the accuracy of the English-Chinese sense links by assessing manually for each example whether the Chinese translation could be substituted to the English verb in a translation of the whole sentence into Chinese. To simplify the task, we reduced the data in the following two ways:

- Only one Chinese word was used as a translation of a synset (for example 鼓动, 挑起, and 煽动all map to 02585050-v glossed as "try to stir up public opinion", according to COW), initially randomly selected (we selected only 鼓动, when Babelfy disambiguated a given verb use as 02585050-v).
- Redundant examples were removed from the evaluation on the grounds that because they are all examples of the same pattern, the validity of one translation should be valid for all other examples.

The results show that 1,598 (4,872 over the whole set of examples) examples were correct, and 2,079 were wrong (5,920). This covers 743 verbs and 869 Chinese words for 1,468 PDEV patterns/senses and 959 WN synsets.

The benefits of this method are to get a fine-grained evaluation of sense links between Chinese words and WordNet senses based on examples. Errors can either be explained by a wrong mapping in COW, but most realistically, the experience of the Chinese annotator is that Chinese translations in COW are context-insensitive, and are only wrong in that sense. This is generally a consequence of the concept of synset which groups words sharing similar meanings, but where members of the synset cannot strictly be substituted in every context (there are no exact synonyms in natural languages).

An example-based approach provides the missing piece of the puzzle. Because examples are linked to patterns, we can also transfer the semantic structures (arguments) from English to Chinese, in order to draft automatically entries for Chinese words as part of a multilingual pattern dictionary. For example, 鼓动was correctly found to link to the second pattern of *agitate*, and we can therefore suggest that when this Chinese verb has [[Anything]] as subject and either [[Human]], [[Institution]] or [[Animal]] as direct object, it means "[[Anything]] makes [[Human | Institution | Animal]] feel anxious, alarmed, or nervous"as in "The Admiralty was sorely agitated by the shipwrights' custom of taking 'chips'."

Last but not least, this method also allows to collect more than one WN sense for a given pattern sense. Thus whenever two patterns point to the same synset, it entails that they are semantically similar, and that PDEV is making a distinction where WordNet isn't, and vice versa. Thus, both patterns of verb *fidget* map to the same WN synset 02058448-v "move restlessly", but PDEV makes a distinction between fidgeting with a [[Physical Object]] and the intransitive use. This method also enables to harvest similarities between patterns belonging to different verbs such as *cooling* and *chilling* in the spirit of WN synsets.

6.3 Ukrainian manual study

We attempted to use the Babelfy disambiguation system for Ukrainian. However, Ukrainian is a less resourced language, and Babelfy returned very few hits, probably because of the limited success or availability of tokenization, part of speech tagging tools, as well as the low coverage of existing lexical resources for Ukrainian. We submitted 20 sentences and only two Ukrainian verbs returned results, but were translated to nouns.

However, we proceeded to evaluate whether parallel resources could reliably be used by lexicographers to automatically draft bilingual dictionaries, and in our case, to align PDEV to Ukrainian. We used the SketchEngine (Kilgarriff et al., 2014) to extract verbs from the OPUS aligned corpus and presented the lexicographer with the Ukrainian word and the sentence pair. The lexicographer's task was to identify the word in the English sentence which translated the Ukrainian verb, if any, and look up in PDEV if a pattern number could be matched.

The evaluation revealed that, out of 100 examples, 36 were problematic:

- 17 cases were pre-processing issues where no English sentence was presented to the user. The lexicographer translated and aligned them to PDEV but could not evaluate the English alignment.
- 9 verbs did not have a direct equivalent in the English translation.
- 6 verbs had problematic English translations, including not appropriate, bad, or incorrect translations. These were corrected and mapped to PDEV.

Thus 64% of examples could be used to link Ukrainian with English. However, only 17% of examples (10% without human intervention) matched an existing PDEV entry, which accounts for 63 verbs not being described yet in PDEV. An example of a satisfactory link is illustrated in examples (2a) and (2b).

(2a) Якщо буде позначено цей пункт , К3б не буде **висувати** лоток з носієм відразу після завершення запису .

(2b) If this option is checked K3b will not **eject** the medium once the burn process finishes .

The pattern illustrated is eject 4 (see Table 3).

Pattern [[Machine]] ejects [[Artifact]]
Implicatures [[Machine]] pushes out [[Artifact]] *This is generally a case of a disc or other hardware being ejected by a computer or other technological device*

Table 3: PDEV pattern 4 of eject

7 Conclusion and Future Work

This paper reports on the evaluation of current linked data solutions to build a multilingual network, which integrates lexicons, ontologies, and corpora to serve Language Learning applications, especially in the process of building learning materials and activities. The paper proposes a data model for the network, in which knowledge can be conveyed from one resource to another, from one language to another. This is particularly useful for language learning, as several views on the data can be created for different audiences or different language topics (meaning, grammar, spelling, etc.). This paper focuses on sense linking for multilingual resources (English, Chinese, and Ukrainian) and proposes several methods to achieve this goal, depending on available resources. Because quality is an essential feature of such an application, the paper runs several evaluations of existing resources and state-of-the-art NLP and WSD systems. The evaluations are quite pessimistic as sense linking success is hindered by errors introduced at various stages, or insufficient coverage of lexical resources.

Extracting reliable links, however, is a major issue in Linguistic Linked Data, and there are various other methods than the ones presented in this paper to achieve it. We are particularly interested in evaluating distributional thesauri automatically constructed from corpora to identify sense candidates, as well as semi-supervised methods, where a few translated examples are provided as seeds to a bootstrapping algorithm.

Perspectives also include evaluating PDEV pattern transfers to languages such as Ukrainian and Chinese, and particularly enable an evaluation of cross-lingual verb semantic preferences. With a view on the language learning application, we intend to evaluate how images and other media can be collected and sense-linked to our network, much like what BabelNet proposes.

References

Eneko Agirre and Aitor Soroa. 2009. Personalizing pagerank for word sense disambiguation. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 33–41. Association for Computational Linguistics.

Enrique Amigó, Julio Gonzalo, Javier Artiles, and Felisa Verdejo. 2009. A comparison of extrinsic clustering evaluation metrics based on formal constraints. *Information Retrieval*, 12(4):461–486.

Amit Bagga and Breck Baldwin. 1999. Cross-document event coreference: Annotations, experiments, and observations. In *Proceedings of the Workshop on Coreference and its Applications*, pages 1–8.

Vìt Baisa, Jane Bradbury, Silvie Cinková, Ismaïl El Maarouf, Patrick Hanks, Adam Kilgarriff, and Octavian Popescu. 2015. Semeval-2015 task 15: A cpa dictionary-entry-building task. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, Denver, Co, USA.

Satanjeev Banerjee and Ted Pedersen. 2002. An adapted lesk algorithm for word sense disambiguation using wordnet. In *Computational linguistics and intelligent text processing*, pages 136–145. Springer.

Francis Bond and Ryan Foster. 2013. Linking and extending an open multilingual wordnet. In *ACL (1)*, pages 1352–1362.

Francis Bond and Kyonghee Paik. 2012. A survey of wordnets and their licenses. In *Proceedings of the 6th International Global WordNet Conference*, pages 64–71.

Lou Burnard. 2007. Reference guide for the british national corpus (xml edition), 2007. *URL http://www. natcorp. ox. ac. uk/XMLedition/URG*.

Antonio Miguel Corbí Bellot, Mikel L Forcada Zubizarreta, Sergio Ortiz Rojas, Juan Antonio Pérez Ortiz, Gema Ramírez Sánchez, Felipe Sánchez Martínez, Iñaki Alegría Loinaz, Aingeru Mayor Martínez, Kepa Sarasola Gabiola, et al. 2005. An open-source shallow-transfer machine translation engine for the romance languages of spain. In *European Association for Machine Translation*.

Ismail El Maarouf, Jane Bradbury, and Patrick Hanks. 2014. Pdev-lemon: a linked data implementation of the pattern dictionary of english verbs based on the lemon model. In *Proceedings of the 3rd Workshop on Linked Data in Linguistics (LDL): Multilingual Knowledge Resources and Natural Language Processing at the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland.

Charles J. Fillmore. 1985. Frames and the semantics of understanding. *Quaderni di Semantica*, 6(2):222–254.

Jorge Gracia, Elena Montiel-Ponsoda, Daniel Vila-Suero, and Guadalupe Aguado-de-Cea. 2014. Enabling language resources to expose translations as linked data on the web. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014), Reykjavik, Iceland, May 26-31, 2014.*, pages 409–413.

Iryna Gurevych, Judith Eckle-Kohler, Silvana Hartmann, Michael Matuschek, Christian M Meyer, and Christian Wirth. 2012. Uby: A large-scale unified lexical-semantic resource based on lmf. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 580–590. Association for Computational Linguistics.

Patrick Hanks and James Pustejovsky. 2005. A pattern dictionary for natural language processing. *Revue Française de linguistique appliquée*, 10:2.

Patrick Hanks. 2012. How people use words to make meanings: Semantic types meet valencies. In A. Boulton and J. Thomas, editors, *Input, Process and Product: Developments in Teaching and Language Corpora*, pages 54–69. Masaryk University Press, Brno.

Patrick Hanks. 2013. *Lexical Analysis: Norms and Exploitations*. MIT Press, Cambridge, MA.

Sebastian Hellmann, Jens Lehmann, Sören Auer, and Martin Brümmer. 2013. Integrating nlp using linked data. In *The Semantic Web–ISWC 2013*, pages 98–113. Springer.

Adam Kilgarriff and Martha Palmer. 2000. Introduction to the special issue on senseval. *Computers and the Humanities*, 34:1–2.

Adam Kilgarriff, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý, and Vít Suchomel. 2014. The sketch engine: ten years on. *Lexicography*, 1(1):7–36.

Michael Lesk. 1986. 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*, SIGDOC '86, pages 24–26, New York, NY, USA. ACM.

Michael Matuschek and Iryna Gurevych. 2013. Dijkstra-wsa: A graph-based approach to word sense alignment. *Transactions of the Association for Computational Linguistics*, 1:151–164.

John McCrae, Dennis Spohr, and Philipp Cimiano. 2011. Linking lexical resources and ontologies on the semantic web with lemon. In *Proceedings of the 8th Extended Semantic Web Conference on The Semantic Web: Research and Applications - Volume*

Part I, ESWC'11, pages 245–259, Berlin, Heidelberg. Springer-Verlag.

Rada Mihalcea. 2005. Unsupervised large-vocabulary word sense disambiguation with graph-based algorithms for sequence data labeling. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, pages 411–418, Stroudsburg, PA, USA. Association for Computational Linguistics.

Andrea Moro, Alessandro Raganato, and Roberto Navigli. 2014. Entity linking meets word sense disambiguation: A unified approach. *Transactions of the Association for Computational Linguistics*, 2:231–244.

Roberto Navigli and Simone Paolo Ponzetto. 2012. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

Roberto Navigli. 2009. Word sense disambiguation: A survey. *ACM Comput. Surv.*, 41(2):10:1–10:69, February.

Elisabeth Niemann and Iryna Gurevych. 2011. The people's web meets linguistic knowledge: Automatic sense alignment of wikipedia and wordnet. In *Proceedings of the 9th International Conference on Computational Semantics*, pages 205–214, Singapore, January.

Darius Pfitzner, Richard Leibbrandt, and David Powers. 2009. Characterization and evaluation of similarity measures for pairs of clusterings. *Knowledge Information Systems*, 19(3):361–394.

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

Liling Tan. 2014. Pywsd: Python implementations of word sense disambiguation (wsd) technologies [software]. https://github.com/alvations/pywsd.

Jörg Tiedemann. 2009. News from opus-a collection of multilingual parallel corpora with tools and interfaces. In *Recent advances in natural language processing*, volume 5, pages 237–248.

Shan Wang and Francis Bond. 2013. Building the chinese open wordnet (cow): Starting from core synsets. In *Sixth International Joint Conference on Natural Language Processing*, page 10. Citeseer.

Renjie Xu, Zhiqiang Gao, Yingji Pan, Yuzhong Qu, and Zhisheng Huang. 2008. An integrated approach for automatic construction of bilingual chinese-english wordnet. In *Proceedings of the 3rd Asian Semantic Web Conference on The Semantic Web*, ASWC '08, pages 302–314, Berlin, Heidelberg. Springer-Verlag.