Navigating Sense-Aligned Lexical-Semantic Resources: THE WEB INTERFACE TO UBY

Iryna Gurevych[1,2], **Michael Matuschek**[1], **Tri-Duc Nghiem**[1],
Judith Eckle-Kohler[1], **Silvana Hartmann**[1], **Christian M. Meyer**[1]
[1]Ubiquitous Knowledge Processing Lab (UKP-TUDA)
Department of Computer Science, Technische Universität Darmstadt
[2]Ubiquitous Knowledge Processing Lab (UKP-DIPF)
German Institute for Educational Research and Educational Information
http://www.ukp.tu-darmstadt.de

Abstract

In this paper, we present the Web interface to UBY, a large-scale lexical resource based on the Lexical Markup Framework (LMF). UBY contains interoperable versions of nine resources in two languages. The interface allows to conveniently examine and navigate the encoded information in UBY across resource boundaries. Its main contributions are twofold: 1) The visual view allows to examine the sense clusters for a lemma induced by alignments between different resources at the level of word senses. 2) The textual view uniformly presents senses from different resources in detail and offers the possibility to directly compare them in a parallel view. The Web interface is freely available on our website[1].

1 Introduction

Lexical-semantic resources (LSRs) are the foundation of many Natural Language Processing (NLP) tasks. Recently, the limited coverage of LSRs has led to a number of independent efforts to align existing LSRs at the word sense level.

However, it is very inconvenient to explore the resulting sense-aligned LSRs, because there are no APIs or user interfaces (UIs) and the data formats are heterogeneous. Yet, easy access to sense-aligned LSRs would be crucial for their acceptance and use in NLP, as researchers face the problem of determining the added value of sense-aligned LSRs for particular tasks.

In this paper, we address these issues by presenting UBY-UI, an easy-to-use Web-based UI to the large sense-aligned LSR UBY (Gurevych et al., 2012). UBY is represented in compliance with the ISO standard LMF (Francopoulo et al., 2006) and currently contains interoperable versions of nine heterogeneous LSRs in two languages, as well as pairwise sense alignments for a subset of them: English WordNet (WN), Wiktionary (WKT-en), Wikipedia (WP-en), FrameNet (FN), and VerbNet (VN); German Wiktionary (WKT-de), Wikipedia (WP-de), and GermaNet (GN), and the English and German entries of OmegaWiki (OW-en/de).

The novel aspects of our interface can be summarized as 1) *A graph-based visualization of sense alignments between the LSRs integrated in* UBY. Different senses of the same lemma which are aligned across LSRs are grouped. This allows intuitively exploring and assessing the individual senses across resource boundaries. 2) *A textual view for uniformly examining lexical information in detail.* For a given lemma, all senses available in UBY can be retrieved and the information attached to them can be inspected in detail. Additionally, this view offers to compare any two senses in a detailed contrasting view.

2 Related Work

Single Resource Interfaces. Web interfaces have been traditionally used for electronic dictionaries, such as the Oxford Dictionary of English. Lew (2011) reviews the interfaces of the most prominent English dictionaries. These interfaces have also largely influenced the development of Web interfaces for LSRs, such as the ones for WN, FN, WKT, or the recently presented DANTE (Kil-

[1]https://uby.ukp.informatik.tu-darmstadt.de

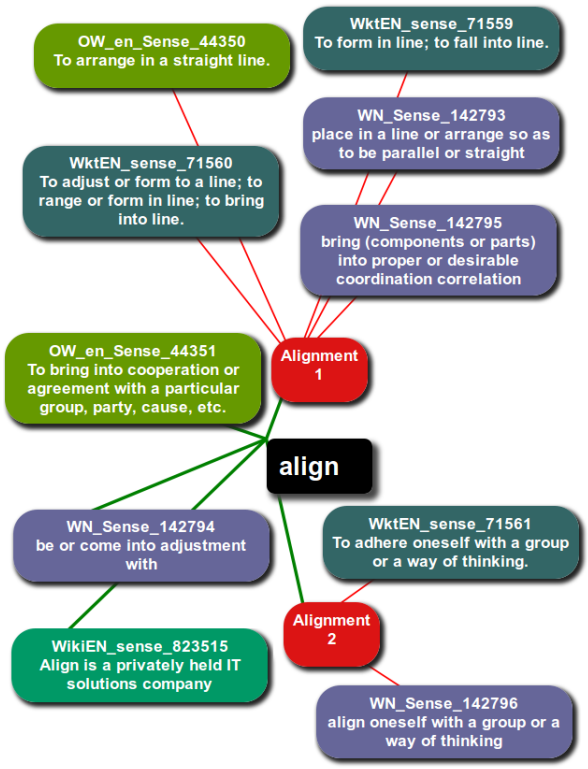

Figure 1: Search result for the verb *align* in the visual view. The aligned senses are connected by sense alignment nodes. Nodes are coloured by resource.

garriff, 2010) which directly adapted the dictionary interface models. All of these Web interfaces have been designed in strict adherence to a specific, single LSR. The UBY-UI is, in contrast, designed for multiple heterogeneous LSRs.

Multi Resource Interfaces. Only a few other Web interfaces are able to display information from multiple LSRs. The majority of them is limited to show preformatted lexical entries one after another without interconnecting them. Popular examples are Dictionary.com[2] and TheFreeDictionary[3]. Similarly, the DWDS interface (Klein and Geyken, 2010) displays its entries in small rearrangable boxes. The Wörterbuchnetz (Burch and Rapp, 2007) is an example of a Web interface that connects its entries by hyperlinks – however, only at the level of lemmas and not word senses.

In contrast, UBY-UI provides hyperlinks to navigate between different *word senses*, as UBY provides mono- and cross-lingual sense alignments between its LSRs. Additionally, UBY-UI supports the direct comparison of two arbitrary word senses present in UBY. In the other multi resource interfaces, this is only possible for whole lexical entries.

Graph-based Interfaces. Two examples for visualizing WN are Visuwords[4] and WordNet explorer[5] that allow browsing the WN synset structure. An example for a cross-lingual graph-based interface is VisualThesaurus[6] which shows related words in six different languages. UBY-UI provides a similar graph-based interface, but combines the information from multiple types of LSRs interlinked by means of sense alignments.

3 UBY – A Sense-Aligned LSR

LMF Model. The large-scale multilingual resource UBY holds standardized and hence interoperable versions of the nine LSRs previously listed. UBY is represented according to the UBY-LMF lexicon model (Eckle-Kohler et al., 2012), an instantiation and extension of the meta lexicon model defined by LMF. Developing a lexicon model such as UBY-LMF involves first selecting appropriate classes (e.g. `LexicalEntry`) from the LMF packages, second defining attributes for these classes (e.g. `part of speech`), and third linking the attributes and other linguisitc terms (such as attribute values) to ISOCat.[7] UBY-LMF is capable of representing a wide range of information types from heterogeneous LSRs, including both expert-constructed resources and collaboratively constructed resources. Representing them according to the class structure of UBY-LMF makes them structurally interoperable. The linking of linguistic terms with their meaning as defined in ISOCat contributes to semantic interoperability.

Sense Alignments. UBY-LMF models a `LexicalResource` as consisting of one `Lexicon` per integrated resource. These `Lexicon` instances can be aligned at the sense level by linking pairs of senses or synsets using

[2]http://www.dictionary.com
[3]http://www.thefreedictionary.com
[4]http://www.visuwords.com
[5]http://faculty.uoit.ca/collins/research/wnVis.html
[6]http://www.visualthesaurus.com
[7]http://www.isocat.org/, the implementation of the ISO 12620 Data Category Registry (Broeder et al., 2010).

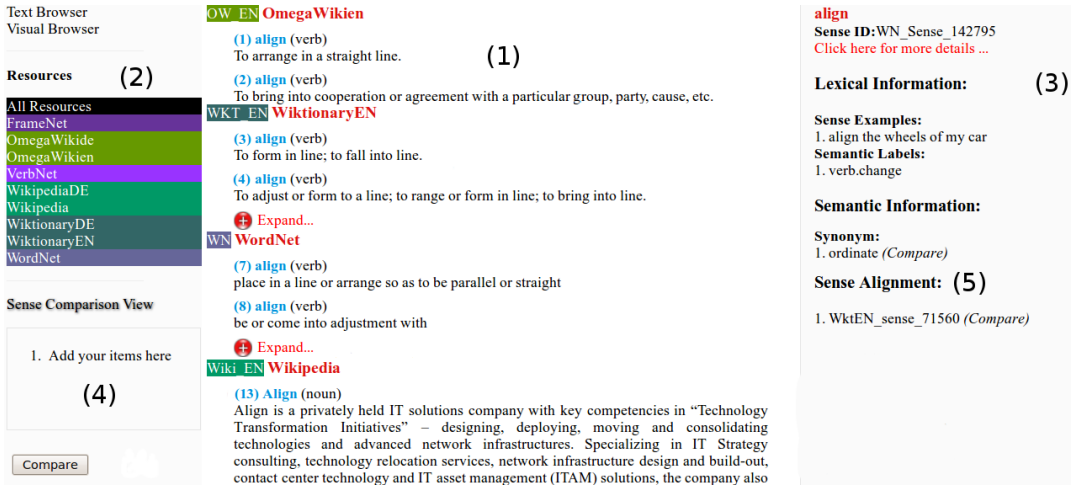

Figure 2: The textual view: (1) Senses of *align*, grouped by resource. (2) Area for selecting resources. (3) Detail view for a selected sense. (4) Drag & drop area for sense comparison. (5) Links to other senses.

instances of the `SenseAxis` class. The resource UBY features pairwise sense alignments between a subset of LSRs. Both monolingual and cross-lingual alignments are present in UBY: WN–WP-en (Niemann and Gurevych, 2011), WN–WKT-en (Meyer and Gurevych, 2011), VN–WN (Kipper et al., 2006), VN–FN (Palmer, 2009), OW-de–WN (Gurevych et al., 2012) and OW–WP, OW-en–OW-de, WP-en–WP-de which are part of the original resources.

The WN–WP-en, WN–WKT-en and OW-de–WN alignments have been automatically created. Please refer to the papers mentioned above for details on the alignment algorithm and detailed statistics.

UBY 1.0 UBY currently contains more than 4.5 million lexical entries, 4.9 million senses, 5.4 million semantic relations between senses and more than 700,000 alignments between senses. There are 890,000 unique German and 3.1 million unique English lemma-POS combinations[8].

4 UBY Web Interface

Technical Basis. UBY is deployed in an SQL database via hibernate, which is also the foundation of the UBY-API. This allows to easily query all information entities within UBY. More details on the UBY-API can be found on the UBY website[9]. The frontend of the Web application is based on Apache Wicket[10].

Visual View. The natural entry point to the visual view is the search box for a lemma[11], and the result is a graph, with the query lemma as the central node and the retrieved senses as nodes attached to it (see figure 1). The sense nodes are coloured according to the source LSRs. To keep the view compact, the definition is only shown when a node is clicked.

The sense alignments between LSRs available in UBY are represented by *alignment nodes*, which are displayed as hubs connecting aligned senses. For generating the alignment nodes, we cluster senses based on their pairwise aligments and include all senses which are directly or transitively aligned. Thus, the visual view provides a visualization of which and how many senses from different LSRs are aligned in UBY. In Figure 1, we show the grouping of senses for the verb *align*. If a user wants to inspect a specific sense in more detail, a click on the link within a sense node opens the textual view described below.

Textual View. While the query mechanism for the textual view is the same as for the visual view, in this case the interface returns a list of senses (see (1) in Figure 2), including definitions, available for this lemma either in all LSRs, or only

[8] Note that for homonyms there may be more than one `LexicalEntry` for a lemma-POS combination.

[9] http://www.ukp.tu-darmstadt.de/uby

[10] http://wicket.apache.org/

[11] Filtering by POS is to be included in a future release.

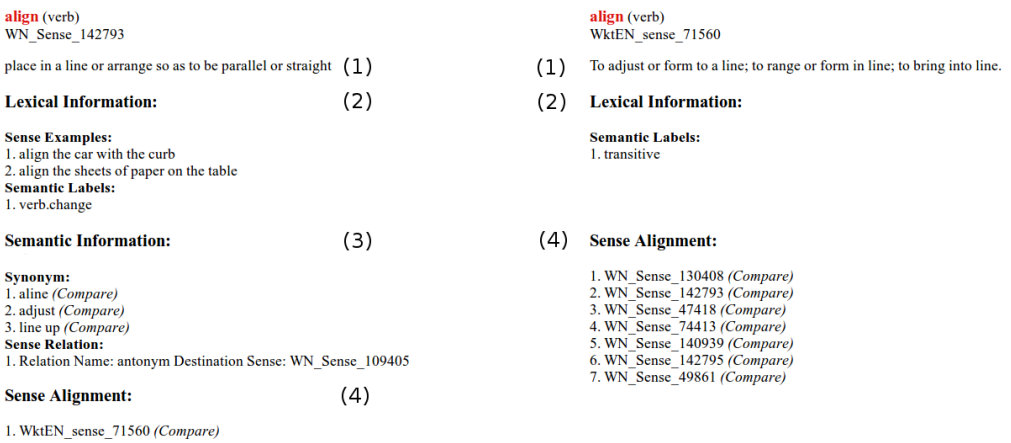

Figure 3: In the sense comparison view, detailed information for two arbitrary senses can be inspected. Below the definition for each sense (1), lexical (2) and semantic (3) information is listed if available. Note the alignment sections (4) which contain links to the aligned senses, as well as links to compare two senses immediately.

those selected by the user (2). Additionally, the LSRs are colour-coded like in the visual view.

For further exploring the information attached to a single sense, clicking on it opens an expanded view on the right-hand side (3) showing more detailed information (e.g. sense examples). Optionally, a full screen view can be opened which allows the user to explore even more information. In the detailed view of a sense, it is also possible to navigate to other senses by following the hyperlinks, e.g. for following sense alignments across LSRs (5).

For comparing the information attached to two senses in parallel, we integrated the option to open a comparison view. For this, the user can directly drag and drop two senses to a designated area of the UI to compare them (4), or click the *Compare* link in the sense detail view (5).

The advantage of the comparison view is illustrated in Figure 3: As the information is presented in a uniform way (due to the standard-compliant representation of UBY), a user can easily compare the information available from different LSRs without having to use different tools, terminologies, and UIs. In particular, for senses that are aligned across LSRs, the user can immediately detect complementary information, e.g., if a WKT sense does not have sense examples but the aligned WN sense does, this additional information becomes directly accessible. To our knowledge, such a contrasting view of two *word senses* has not been offered by any resource or UI so far.

5 Conclusions and Future Work

In this paper, we presented a Web interface to UBY, a large-scale sense-aligned LSR integrating knowledge from heterogeneous sources in two languages. The interface combines a novel, intuitively understandable graph view for sense clusters with a textual browser that allows to explore the offered information in greater detail, with the option of comparing senses from different resources.

In future work, we plan to extend the UI to allow editing of the alignment information by the users. The rationale behind this are the errors resulting from automatic alignment. A convenient editing interface will thus help to improve the underlying resource UBY. Another goal is to enhance the display of alignments across multiple resources. Right now, we use pairwise alignments between resources to create sense clusters, but as we plan to add more sense alignments to UBY in the future, the appropriate resolution of invalid alignments will become necessary.

Acknowledgments

This work has been supported by the Volkswagen Foundation as part of the Lichtenberg-Professorship Program under grant No. I/82806. We thank Richard Eckart de Castilho and Zijad Maksuti for their contributions to this work.

References

Daan Broeder, Marc Kemps-Snijders, Dieter Van Uytvanck, Menzo Windhouwer, Peter Withers, Peter Wittenburg, and Claus Zinn. 2010. A Data Category Registry- and Component-based Metadata Framework. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC)*, pages 43–47, Valletta, Malta.

Thomas Burch and Andrea Rapp. 2007. Das Wörterbuch-Netz: Verfahren - Methoden - Perspektiven. In *Geschichte im Netz: Praxis, Chancen, Visionen. Beiträge der Tagung .hist 2006*, Historisches Forum 10, Teilband I, pages 607–627. Berlin: Humboldt-Universität zu Berlin.

Judith Eckle-Kohler, Iryna Gurevych, Silvana Hartmann, Michael Matuschek, and Christian M. Meyer. 2012. UBY-LMF - A Uniform Model for Standardizing Heterogeneous Lexical-Semantic Resources in ISO-LMF. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC)*, pages 275–282, Istanbul, Turkey.

Gil Francopoulo, Nuria Bel, Monte George, Nicoletta Calzolari, Monica Monachini, Mandy Pet, and Claudia Soria. 2006. Lexical Markup Framework (LMF). In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC)*, pages 233–236, Genoa, Italy.

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 (EACL)*, pages 580–590.

Adam Kilgarriff. 2010. A Detailed, Accurate, Extensive, Available English Lexical Database. In *Proceedings of the NAACL-HLT 2010 Demonstration Session*, pages 21–24, Los Angeles, CA, USA.

Karin Kipper, Anna Korhonen, Neville Ryant, and Martha Palmer. 2006. Extending VerbNet with Novel Verb Classes. In *Fifth International Conference on Language Resources and Evaluation (LREC)*, pages 1027–1032, Genoa, Italy.

Wolfgang Klein and Alexander Geyken. 2010. Das Digitale Wörterbuch der Deutschen Sprache (DWDS). *Lexicographica*, 26:79–96.

Robert Lew. 2011. Online dictionaries of English. In Pedro A. Fuertes-Olivera and Henning Bergenholtz, editors, *E-Lexicography: The Internet, Digital Initiatives and Lexicography*, pages 230–250. London/New York: Continuum.

Christian M. Meyer and Iryna Gurevych. 2011. What Psycholinguists Know About Chemistry: Aligning Wiktionary and WordNet for Increased Domain Coverage. In *Proceedings of the 5th International Joint Conference on Natural Language Processing (IJCNLP)*, pages 883–892, Chiang Mai, Thailand.

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 (IWCS)*, pages 205–214, Oxford, UK.

Martha Palmer. 2009. Semlink: Linking PropBank, VerbNet and FrameNet. In *Proceedings of the Generative Lexicon Conference (GenLex)*, pages 9–15, Pisa, Italy.