Kamusi Pre:D - Source-Side Disambiguation and a Sense Aligned Multilingual Lexicon

Martin Benjamin
Distributed Information
Systems Laboratory (LSIR)
École Polytechnique
Fédérale de Lausanne
martin.benjamin@epfl.ch

Amar Mukunda
Distributed Information
Systems Laboratory (LSIR)
École Polytechnique
Fédérale de Lausanne
amar.mukunda@epfl.ch

Jeff Allen
Products & Innovation: User
Assistance, Language
Management & Translation
SAP France
jeff.allen@sap.com

Abstract

This paper discusses Kamusi Pre:D, a system to improve translation by disambiguating word senses in a source document with reference to a large concept-based lexicon that is aligned by sense across numerous languages. Currently under active development, the program prompts users to select the intended meaning when polysemous terms occur, and gives the user the option to select multiword expressions instead of individual words when the MWE occurs as a lexicalized dictionary entry. The disambiguated text is then automatically matched to sense-specific translation equivalents that have been aligned across languages. Pre:D is intended to integrate with existing translation tools, but greatly improve accuracy by involving human intelligence in vocabulary selection, both through manual document review of ambiguous terms and by reference to the underlying curated multilingual Kamusi dictionary data. Pre:D will aid accurate vocabulary translation among a wide range of language pairs, most currently unserved, and offer significant advantages in time, effort, and quality for multilingual translation projects by disambiguating a document one time for concepts that can be rendered appropriately across numerous languages.

1 Introduction

Kamusi Pre:D offers a new approach to translation by disambiguating word senses on the source side that are matched to human-confirmed vocabulary equivalents in any target language. As a fine-grained knowledge-based system (Ponzetto and Navigli 2010), Pre:D has the potential for much greater term accuracy than algorithmic word sense disambiguation (WSD) (Vickrey et al. 2005, Agirre and Edmonds 2006) or magic wand machine translation (MT) approaches (Chan et al. 2007); while the program will evolve to employ statistics and machine learning (Tyers et al. 2012) in ranking senses for recommendation, it is the informed interaction between person and machine in selecting meanings that will enable concepts to be pinpointed across languages. Predisambiguation will be especially relevant for the vast majority of language pairs for which no parallel text corpora exist to even attempt rudimentary statistical translation (e.g. Ng et al. 2003, Specia et al. 2005), but this tool for manual preparation is expected to improve quality substantially even for well-trod language pairs. In the Kamusi Pre:D interface, documents in any project source language are matched against the multilingual Kamusi Project lexicon. Terms that have multiple senses in the Kamusi dataset are highlighted in the source document, much as misspellings are in a spellchecker. When the user hovers over a term flagged as ambiguous, the various sense definitions are displayed. After the user selects the intended meaning, known equivalent terms for any target language are passed to computer assisted translation (CAT) or MT, where

Proceedings of the 37th Conference Translating and the Computer, pages 27–32,
London, UK, November 26-27, 2015. ©2015 AsLing

statistics and rules can be brought to bear with the sense-restricted vocabulary (Eisele et al. 2008).

Kamusi Pre:D has three anticipated use cases:
1) Immediate hand translation. In this case, the user can drill personally to the translation level, selecting a matched equivalent as part of the review process.
2) Preparation for a translation team. In this case, an initial user tags the senses in the source language, and the options for matched equivalents are presented to individual translators for each target language.
3) Preparation for machine translation. In this case, the user tags senses in the source language, and the options for matched equivalents are selected by MT for each target language.

2 Individual Words

The Kamusi lexicon is an expanding resource that is working toward monolingual sense-disambiguated dictionaries for each language, with parallel or similar concepts marked and linked across languages to create a multilingual semantic matrix. In 2015 the project brought in more than 1.2 million terms in over 20 languages (with numbers growing steadily), aligned by concept. These terms, from the Open Multilingual Wordnet (discussed as a basis for WSD by Navigli 2006) and other sources, currently only in canonical form, have not yet been subject to the human review features Kamusi has developed for dictionary-quality entries. For example, the Wordnet import contains many erroneous translation equivalents such as French “lumière” for the low calorie sense of “light”, which should be fixed by Kamusi participants through pending crowdsourcing features. However, the provisional data proves the concept that sense-specific vocabulary can be identified for Pre:D in any language for which data has been linked.

Word forms are stored in Kamusi as data elements associated with a specific sense of a lemma. That is, inflections such as “saw” map to the different instances of “see” within the database, so an occurrence of “saw” in a document will find the various senses of that verb in Kamusi in addition to the “saw” that cuts wood. (Pre:D will embed part-of-speech tagging as early future work, after evaluating which existing off-the-shelf tagger will best serve multilingual expansibility.) The Kamusi structure is designed so that inflected forms can be linked across languages, but getting the data paired at that level will be a lengthy process; until the data meets the design, Pre:D can only identify canonical vocabulary matches for the inflected forms that are contained in the dataset, and pass the task of target-side grammatical transformations to human or machine processes. Moreover, language-specific rule-based parsing algorithms are necessary to identify lemmatic forms in some languages, such as the rules Kamusi developed for dictionary users to find the verb stem from the tens of millions of potential forms of each Swahili verb.

3 Multiple Words

Within the data design, multi-word expressions are treated as lexicalizable concepts. Identifying MWEs is a fraught topic for natural language processing (Carpuat and Wu 2007, Carpuat and Diab 2010), for which a cross-lingual concept-based data reference can prove particularly beneficial. The general principle for Kamusi is that an MWE should be a dictionary entry if its meaning cannot be determined by the individual entries for the sum of its parts, with a preference to include entries for borderline cases such as “break water” during childbirth.[1] Having a monolingual dictionary entry provides the opportunity to diagram translation equivalents for the concept in any other language. MWEs in Kamusi can be marked to show the point of potential separability, such as "drive || up the wall". Furthermore, because MWEs are treated as normal dictionary entries with POS, their inflected forms should ordinarily be included, e.g. "*driving* || up the wall".

The first round of Pre:D programming will identify contiguous MWEs that exist in the lexicon. For example, the term "African fish eagle" will locate the various terms for the polysemous parts, and highlight that the combination also forms a known MWE. The user will then be able to select whether each word should be treated in its own right, or whether the unit for translation is the full expression. In cases where a word or words could be part of overlapping MWEs, Pre:D will present the full slate of options.

A first complexity for subsequent programming will be to correctly handle separated expressions (Simard et al. 2005). For example, "drive || up the wall" could hypothetically be separated by a lengthy list of annoyed people. Pre:D will find that "drive" in the database can be followed by separated elements, and therefore continue scanning the sentence for eligible follow-on parts. If "up the wall" is located, the unity will be highlighted for the user to confirm as the intended term, and to select its meaning if there are multiple options. The expression will be marked at point of first contact, i.e. "the noise drives everyone up the wall" will be handed off as "the noise {drives up the wall} everyone".

A second complexity will be rule-based expressions (Ahsan et al. 2010), such as those created with auxiliary verbs. In English, for example, Kamusi verb entries contain participles such as "seen" and "seeing", but not constructed inflections such as "had seen" and "is seeing" (much less separated constructions such as "had for many years been seeing"). We will code rules to identify multi-word constructions with conjugated English auxiliary verbs, including separability. These rules are not generalizable, however, and similar efforts will need to be undertaken for other languages in order to properly survey their source documents.

A third complexity for future work is replaceability. Design of Pre:D has pointed to the need for a new field within the MWE framework for Kamusi. In an MWE such as "run up *a* tab", the article can be replaced by a set of pronouns, by named entities ("run up *Bob's* tab"), or by other terms ("run up *the bride's father's* tab"). In response to this need, Kamusi will program a feature for replaceable elements to be marked within dictionary entries in the

[1] In the Oxford English Dictionary (www.oed.com), for example, *break* "to burst" of a bodily purulence, is verb sense 4 after 14 earlier sub-senses and *water* as "amniotic fluid" is noun sense 19. A human reader could technically find both senses and correlate the meaning, but it would be difficult monolingually. The online Larousse Dictionnaire Anglais-Français (www.larousse.fr) has "her waters broke" as an example in line 39 of the result for "water". A statistical approach would be vanishingly unlikely to propose the correct senses of the individual terms, as shown by the failure of all online translation services to correctly render "her water broke" in any tested language other than a single instance of Google Translate suggesting the correct German. Were a machine to correctly identify the combination, through collocation and domain context, the chances of finding a correct translation through parallel text are small to nil; Linguee (www.linguee.com) finds two acceptable translations to French from 27 nearby occurrences of "her", "water", and "broke". The vast majority of languages have no parallel text for corpus analysis, and almost none have enough to train MT on infrequent occurrences. However, querying native speakers (the Kamusi method for collecting data when experts are unavailable) or experts (aided by dictionaries or Sketch Engine) yielded us the preferred equivalent 100% of the time, for languages from Estonian to Swahili. In an electronic dictionary, there is no penalty other than time in erring on the side of caution by including a technically redundant entry. By contrast, human readers gain by finding a clear meaning and deliberate translations for the term as an MWE, and machines have assured data at hand rather than cycling through computations with tenuous results.

same way as separability. (Far-)future work will attempt to denote the set of items that are replaceable for a given expression, using corpus analysis and machine learning to determine, for example, that “take [a] seat” can only be replaced with possessive pronouns, while “drive [someone] crazy” can be replaced with any sentient being. In the early phases, however, Pre:D will be restricted to noting that an MWE has been filled with a replaceable, e.g. “take” will search Kamusi for possible follow-ons, and treat separated elements as a unified translation term if an item occurs with which it is joined in an entry marked for replaceability, such as “[a] seat” or “[a] shower”.

The goal of Pre:D is to analyze documents for the various elements above in combination. For example, “he had fried rice” should notice that “had fried” is a potential inflected multi-word construction deriving from “fry”, that “fried rice” is an MWE in the database, and that there is overlap between the two possible expressions. Some time is needed, however, for the programming to achieve all of its specs for handling the multiple complexities surrounding MWEs.

4 Predictive Aids

Various aspects of intelligence will be built into Pre:D over time. To begin with, the sense choices that a user makes early in a document will be used to raise those same senses as top recommendations each additional time the sense occurs. Further, when the data supports it, users will be able to have the program preference terminology from a selected domain, or eventually benefit from automated domain selection (Buitelaar et al. 2006). At a later stage, Pre:D will be married to current techniques (Costa-jussà and Fonollosa 2015) based on statistics, collocations (McKeown and Radev 1999, Lü and Zhou. 2004), ontologies, and other types of analysis, in conjunction with partners who are working from those computational directions. However, automation is always seen as an aid rather than a goal, with human confirmation of intended meanings on the source side being the key to the computer selecting reliable translation terms.

5 Interactive Growth

Important to the functioning of Kamusi Pre:D is responsiveness to missing entries. Pre:D focuses on one sentence at a time (keeping track of repeat occurrences of a term within a document for weighting later suggestions based on a user’s early selections), presenting each term in a sidebar along with dictionary options displayed by predicted weight. Oftentimes, a term or a sense will be missing in the source language, or a translation equivalent will be missing in other languages. If a user does not find the intended sense of a term among the options, the Pre:D interface will provide a path to submit an entry for the item directly to the project. Alternatively, the user can send the item as a query, with the source sentence transmitted as a contextual example for the production of a new entry. Terms that exist in the source language data but have not been produced in target languages will be given elevated priority in the workflow, with the potential for participants in Kamusi lexicon development to provide reliable vocabulary equivalents for missing vocabulary within a workable timeframe. Kamusi has a crowdsourcing system for members of a speaker community to play games that result in validated language data, to which missing terms will be submitted with a ticking clock and bonus points for rapid responses, and from which results will be incorporated in the larger data set and also transmitted to the original requester. In future development, a system will be implemented to harvest, with permission, completed hand translations for usage examples and translation memory.

Named entities present a special set of challenges that will be addressed as development progresses. Pre:D will integrate code and data from AIDA early on (Yosef et al. 2011). Kamusi will aggregate named entities to the extent possible from open data, and present these terms as word or MWE disambiguation options. However, documents will ordinarily include named entities that are not in existing available datasets, particularly if English is not the source language, or are otherwise ambiguous.[2] Therefore, users will be able to label named entities that are not correctly tagged; these items will be passed to translators as inoperable, and passed to Kamusi for possible inclusion in the named entity data set. Named entities may or may not be translated in the multilingual data, e.g. "Geneva" has numerous translations, whereas "Barack Obama" will be largely consistent for all languages that use the Latin character set. In future work, named entities that are translated on the target side will be returned to the database for validation to be included in the second language.

6 Projections

The first generation of the Pre:D software is intended to be ready for demonstration by late November 2015, with functionality among more than twenty languages. Pipeline features, including full MWE support and push/pull integration with lexicon development, will be added as soon as core features are operational. When complete, Kamusi Pre:D will be ported as a front-end service to provide vocabulary for CAT and MT applications. Individual users will find Pre:D to be an essential tool for accurate vocabulary translation among a wide range of language pairs, most currently unserved, while organizations will recognize significant advantages in time, effort, and quality by disambiguating a document one time for concepts that can be rendered appropriately across numerous languages.

References

Eneko Agirre and Philip Edmonds. 2006. *Word Sense Disambiguation: Algorithms and Applications*, Dordrecht: Springer (Text, speech, and language technology series, edited by Nancy Ide and Jean Véronis, volume 33).

Arafat Ahsan et al. 2010. *Coupling Statistical Machine Translation with Rule-based Transfer and Generation*, AMTA- The Ninth Conference of the Association for Machine Translation in the Americas. Denver, Colorado.

Paul Buitelaar et al. 2006. *Domain Specific WSD*, in Agirre and Edmonds 2006, pp 275-298.

Marine Carpuat and Dekai Wu. 2007. *Improving Statistical Machine Translation using Word Sense Disambiguation*, Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 61–72, Prague, June 2007.

Marine Carpuat and Mona Diab. 2010. *Task-based Evaluation of Multiword Expressions: a Pilot Study in Statistical Machine Translation*, Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL, pages 242–245, Los Angeles, California, June 2010.

Yee Seng Chan, Hwee Tou Ng, and David Chiang. 2007. *Word Sense Disambiguation Improves Statistical Machine Translation*, Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, pages 33–40, Prague, Czech Republic, June 2007.

[2] For example, in an article about Nairobi County, AIDA does not include "Pumwani Division", and correctly identifies "Central Division" as a named entity but suggests it is part of the US National Basketball Association rather than a political area within a city.

Marta Costa-jussà and José Fonollosa. 2015. *Latest trends in hybrid machine translation and its applications*, Computer Speech and Language 32 (2015) 3–10.

Andreas Eisele et al. 2008. *Using Moses to Integrate Multiple Rule-Based Machine Translation Engines into a Hybrid System*, Proceedings of the Third Workshop on Statistical Machine Translation, pages 179–182, Columbus, Ohio, USA, June 2008.

Yajuan Lü and Ming Zhou. 2004. *Collocation Translation Acquisition Using Monolingual Corpora*, Proceedings of ACL 2004 Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics Article No. 167.

Kathleen McKeown and Dragomir Radev. 1999. *Collocations*, In Robert Dale, Hermann Moisl and Harold Somers, (editors), A Handbook of Natural Language Processing. Marcel Dekker, New York.

Roberto Navigli. 2006. *Meaningful Clustering of Sense Helps Boost Word Sense Disambiguation Performance*, Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL, pages 105–112, Sydney, July 2006.

Hwee Tou Ng, Binn Wang, and Yee Seng Chan. 2003. *Exploiting Parallel Texts for Word Sense Disambiguation: An Empirical Study*, ACL 200303 Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1 Pages 455-462.

Simone Ponzetto and Roberto Navigli. 2010. *Knowledge-rich Word Sense Disambiguation Rivaling Supervised Systems*, Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pages 1522–1531, Uppsala, Sweden, 11-16 July 2010.

Michel Simard, et al. 2005. *Translating with non-contiguous phrases*, Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP), pages 755–762, Vancouver, October 2005.

Lucia Specia, Maria das Graças Volpe Nunes and Mark Stevenson. 2005. *Exploiting Parallel Texts to Produce a Multilingual Sense Tagged Corpus for Word Sense Disambiguation*, Recent Advances in Natural Language Processing (RANLP-2005), Borovets, pp. 525-531.

Francis M. Tyers, Felipe Sánchez-Martínez, Mikel L. Forcada. 2012. *Flexible finite-state lexical selection for rule-based machine translation*, Proceedings of the 16th EAMT Conference, 28-30 May 2012, Trento, Italy.

David Vickrey, et al. 2005. *Word-Sense Disambiguation for Machine Translation*, Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language, Processing (HLT/EMNLP), pages 771–778, Vancouver, October 2005.

Mohamed Amir Yosef, et al. 2001. *AIDA: An Online Tool for Accurate Disambiguation of Named Entities in Text and Tables*, in Proceedings of the 37th International Conference on Very Large Databases, VLDB 2011, p. 1450–1453, Seattle, WA.