Kamusi Pre:D – Lexicon-based source-side pre-disambiguation for MT and other text processing applications

Martin Benjamin
École Polytechnique Fédérale de Lausanne
IC LSIR, BC 114, Station 14, CH 1015 Lausanne, Switzerland
Email: martin.benjamin@epfl.ch

Abstract

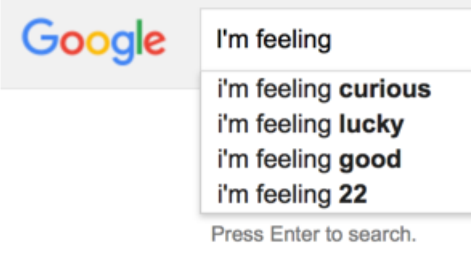
Kamusi has been developing a system to analyze texts on the source side and present users with sense-specified dictionary options. Similarly to spellcheck, the user selects the intended meaning. We then use a multilingual lexical database to bridge to matching vocabulary in other languages. When paired with Freeling, additional pre-processing is possible for several languages. Integration with MT via Moses and Apertium is planned, but not yet undertaken. MWEs treatment is important. An MWE is lexicalized in the Kamusi database and marked for separability, with a definition and translation equivalents (one or more words) in other languages. When the initial term of an MWE appears in the source text, Pre:D queries the database and scans the sentence for all MWEs that could follow. The user can select the relevant MWE rather than the component words. A user can submit a missing sense or MWE for inclusion in the lexicon. Named entities can also be identified from data sources or by users and rendered appropriately across languages. When users agree, we will also use sense-tagged sentences for machine learning. A prototype of the core system is already functional.

Keywords: machine translation; multilingual lexicography; multiword expressions; word sense disambiguation; natural language processing

1. Introduction

Figure 1: Google.com homepage

Do you remember the Google "I'm Feeling Lucky" button? It still appears on the google.com homepage. Start typing your search query, though, and the lottery is replaced with a set of options that try to narrow down your intent. Whether you choose one of those suggestions, or complete your search query, Google will present you with your top search results. Unless you jump through a variety of hoops in the setup menu, automatic machine navigation is dead. As it turns out, people prefer to see their options and make intelligent decisions about the information they seek, versus having a machine make guesses that may well be wrong. You have probably faced a similar battle when your phone autocorrects your words, and may prefer an input method that makes it easy for you to choose a predicted word from a list of candidates, or does not predict for you at all. Hilarious websites chronicle the worst failures of auto-correction[1], whereas spellcheck systems rarely cause comment. Spellcheck and auto-predict both use algorithms that compare input with items in a lexicon, but the former empowers the user with a list of choices. "I'm Feeling Lucky" was retired, and users are distressed by auto-correct, because people prefer the power to determine their intent. This paper discusses a new system, Kamusi Pre:D, that provides a layer between a source document and machine translation, for the user to select their intended meanings from a lexicon in the source language that is matched at a sense level to eligible vocabulary on the target side.



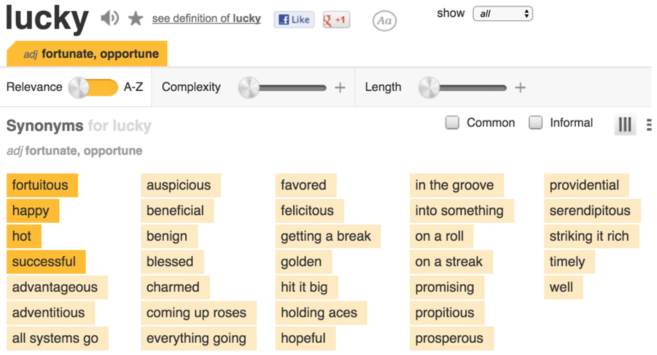
I'm feeling

i'm feeling **curious**
i'm feeling **lucky**
i'm feeling **good**
i'm feeling **22**

Press Enter to search.

Figure 2: Google.com homepage behavior

Part of the role of lexicography is to produce lists of options that help a reader make intelligent decisions. These lists can include inflected forms, alternate spellings, categorical pertinence (ontologies, terminologies), and many more. In monolingual lexicography, a primary task is a list of the different senses that can be represented by a single term – both the disambiguation of polysemy, and possible membership of a term as an element of a longer expression. In bilingual lexicography, the primary task is a list of terms in the second language that convey a meaning close to the source. While scholars can argue at length about whether a definition captures the full essence of a term (Pustejovsky and Rumshisky, 2008; Hanks, 2015), or whether terms in different languages can ever truly share a meaning (Yong and Peng 2007; Tarp, 2008), the average writer or translator often approaches a lexicon with the much more practical goal of selecting the best word for the immediate context.



lucky

adj **fortunate, opportune**

Synonyms for lucky

adj fortunate, opportune

fortuitous
happy
hot
successful
advantageous
adventitious
all systems go
auspicious
beneficial
benign
blessed
charmed
coming up roses
everything going
favored
felicitous
getting a break
golden
hit it big
holding aces
hopeful
in the groove
into something
on a roll
on a streak
promising
propitious
prosperous
providential
serendipitous
striking it rich
timely
well

Figure 3: When users search for the mot juste, *they often seek an uncomplicated list. (http://www.thesaurus.com/browse/lucky)*

2. Predisambiguation tasks for knowledge-based translation

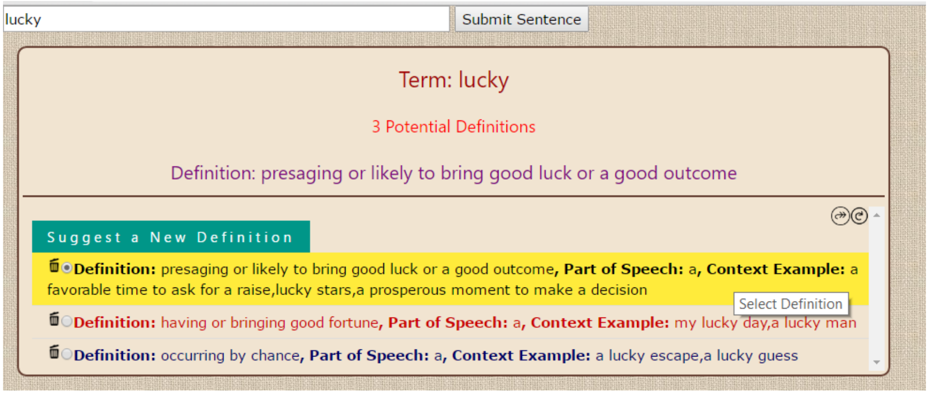
Pre:D[2] combines several elements in the effort to guide translation toward the most appropriate vocabulary option. On the source language side, the program helps the

[1] Enjoy http://www.damnyouautocorrect.com.

[2] Pre:D is short for "predisambiguation": when a term could have more than one meaning, figure out the original intent *before* trying to translate it. The colon is a stylistic flourish.

user identify the various translation items, whether those are single words or lexicalized party terms[3]. Once those items have been determined, the user can choose the sense that most closely matches the intended meaning. Finally, if the item has more than one equivalent in the target language, the user may select among the translation options.

It should be noted that our development effort focuses on the middle steps at the moment. Many of the important challenges for identification of translation items, such as morphological parsing, lemmatization, tokenization, and part of speech identification, are addressed for technologically high-tier languages with packages to be installed at the front end of the Pre:D process.[4] For languages with fewer resources, our data should eventually support identification of inflected forms, at least when those forms can be catalogued in a reasonable list (challenges such as parsing the 900,000,000 valid forms of a given verb in Kinyarwanda, just one of hundreds of agglutinative African languages, are not on the lexicographic agenda, though could be done algorithmically as student research projects using routines similar to the parser we have developed for Swahili); our hope is that we can collaborate with other research teams to expand existing resources like NLTK for currently-excluded languages. Nor do we have in-house expertise in syntactic analysis or integration with MT (machine translation) programs such as Apertium [5]. Our assumption is that the accurate vocabulary output that Pre:D generates can be fed to MT in a way that can be processed into coherent text on the target side; making it so will be future work for graduate students or interested partners.



Term: lucky

3 Potential Definitions

Definition: presaging or likely to bring good luck or a good outcome

● **Definition:** presaging or likely to bring good luck or a good outcome**, Part of Speech:** a**, Context Example:** a favorable time to ask for a raise,lucky stars,a prosperous moment to make a decision

○ **Definition:** having or bringing good fortune**, Part of Speech:** a**, Context Example:** my lucky day,a lucky man

○ **Definition:** occurring by chance**, Part of Speech:** a**, Context Example:** a lucky escape,a lucky guess

Figure 4: Pre:D display of source-side senses [temporary styling]

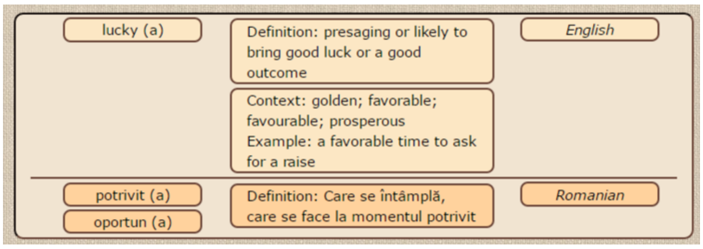
[3] Kamusi is re-branding the term “multiword expression” (MWE) with the more user-friendly “party term” for the benefit of Pre:D users to quickly understand that these are sets of words that play together.

[4] Choosing between, or combining the services of, the Natural Language Toolkit (NLTK, http://www.nltk.org/) and Freeling (http://nlp.lsi.upc.edu/freeling/node/1) is a current task.

[5] https://www.apertium.org

2.1 Core functionality

The central function of Pre:D is to identify source terms and offer closely-matched concepts for translation. The program works one sentence at a time in order to conserve processing resources, though we maintain memory of a user's selections from earlier in the same document in order to prioritize result rankings for recurring terms. In the simple case, the sentence is divided into words, the words are displayed with one or more definitions, and the user clicks the meaning that corresponds, as shown in Figure 4. The vocabulary for the selected concept is then matched to a corresponding term or terms in the target language, as shown in Figure 5, for a human or machine translator to then make syntactically and morphologically coherent.


lucky (a)

Definition: presaging or likely to bring good luck or a good outcome

English

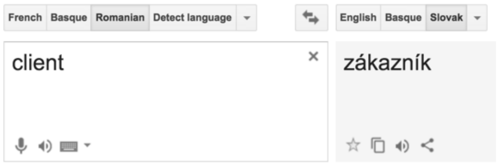
Context: golden; favorable; favourable; prosperous
Example: a favorable time to ask for a raise

potrivit (a)
oportun (a)

Definition: Care se întâmplă, care se face la momentul potrivit

Romanian

Figure 5: Pre:D translation options matched to the source-side sense [temporary styling]



client

zákazník

Figure 6: Google Translate proposal from Romanian to Slovak, returning single option

Vocabulary is matched among every language in the system. Kamusi's forays into new languages begin with open or shared data that is either pre-aligned (e.g., through Wordnet or Wikidata identifiers) (Benjamin, 2016) or matched by game players[6] to a concept in a pivot language, usually English. Auto-magic programs like Google Translate also pass through English, but they make a single guess about the source-side intent and present the user with a computed translation term as a fait accompli – a risky form of "truthiness"[7]. In stark contradistinction, Pre:D attempts to overcome the lack of human confirmation of the link between Language A and Language C by displaying all we know about the chain. Figure 7, for example, shows computed translations between Romanian and Slovak that pass via the Princeton Wordnet (Fellbaum, 1998); the data includes definitions in Romanian and English but not in Slovak, and Kamusi presents the available information for the user to make an informed decision. Until a link between languages has been explicitly validated, the philosophy is to replace absolute truth claims with the evidence of why,

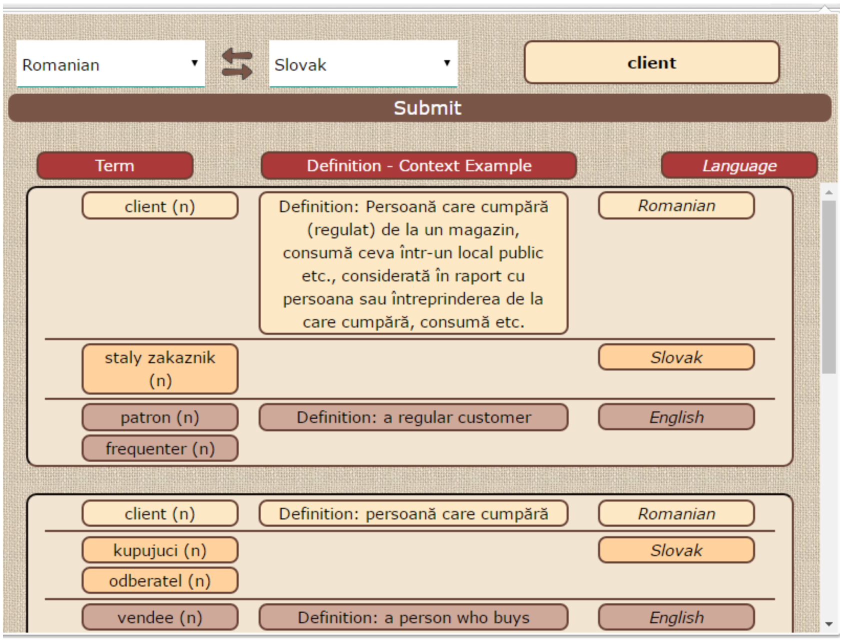
[6] A video demonstrating a version of DUCKS (Data Unified Conceptual Knowledge Sets) is available at https://youtu.be/AK1D7IRUifs . This video shows linking datasets within one language, but the general principle is the same for joining a second language: drag and drop the term on the left to the matching English sense, if any. DUCKS was programmed in May 2016, and has not yet been elsewhere described.

[7] http://www.merriam-webster.com/press-release/2006-word-of-the-year

with rounding errors, a proposed translation achieves “proximity to the truth of language” (Chen, 2016).

2.2 Multiword Expressions

Party terms present a number of challenges for both lexicography and MT (Sag et al, 2002) that Kamusi addresses through data and computational techniques. From a lexicographical perspective, the basic premise is that multi-wordiness is uninteresting – what is important is whether a set of words contains a meaning that cannot be derived from the sum of its parts. Party terms with discrete meanings can be documented as dictionary entries, with definitions for each sense. The concept definitions, in turn, link to the related idea in other languages, regardless of whether those concepts are single words or party terms on the target side. For example, “in a nutshell” maps to “en résumé” in French and “kwa kifupi” in Swahili, leaving aside any thoughts of nuts or shells. Other translation systems, of course, also build repertoires of party terms and known translations, so there is some chance of automatically serving a user correct information based on parallel corpora or translation memory (e.g., Google Translate is correct in both French and Swahili for “in a nutshell”, only correct in French for “it is raining cats and dogs”, and wrong in both for “the spring in her step”). Pre:D, however, introduces several further steps that have not previously been tried. First, when a party term is separable, that separability is marked in the data, e.g. “give [somebody] a break”. The graph database can then identify all the party terms that begin with the word or words before the separability marker, e.g. “give [] a break” and “give [] the time of day”. After converting “gave” to “give”, it is then a simple computational task to scan the remainder of the sentence to see if “a break” or “the time of day” appears downstream, no matter how many words intervene. The n-gram problem that defeats state-of-the-art MT is erased; Pre:D will discover the party terms even in a sentence such as “The attorney general *gave* the CEO, who had manipulated emissions controls, lied to safety inspectors, and defrauded consumers, *a break* when she did not prosecute him.” Finally, the user can select whether a word should be



Romanian ⇆ Slovak

client

Term	Definition - Context Example	*Language*
client (n)	Definition: Persoană care cumpără (regulat) de la un magazin, consumă ceva într-un local public etc., considerată în raport cu persoana sau întreprinderea de la care cumpără, consumă etc.	*Romanian*
staly zakaznik (n)		*Slovak*
patron (n) frequenter (n)	Definition: a regular customer	*English*
client (n)	Definition: persoană care cumpără	*Romanian*
kupujuci (n) odberatel (n)		*Slovak*
vendee (n)	Definition: a person who buys	*English*

Figure 7: Kamusi proposal from Romanian to Slovak, showing all known options

treated independently, or as an element of a party term. In the previous sentence, for example, Pre:D will offer glosses for “attorney”, “general”, and “attorney general”. How to deal with the syntax of separated party terms on the target side once a translation term is identified is a future project.

2.3 Adding and improving senses

Users are encouraged to submit senses that have not yet been documented in Kamusi, both for individual words and for party terms. Party terms can be linked together with buttons on the interface. The feature to add a sense is rudimentary at the moment, but is scheduled to evolve into a suite of lexicographic options for ambitious contributors. The human-computer interaction (HCI) chore is to replace our current complex Edit Engine with a sequence of steps that the user finds simple and manageable, that results in consistent, adequate, and valid data.

When a user encounters an inadequate definition, that sense can be flagged for future review by other Kamusi users. This is important because many definitions from the Princeton Wordnet could bear improvement, e.g., “policewoman: a woman policeman”. A game for improving definitions has been programmed (Benjamin 2015), and awaits activation pending server issues.

2.4 Names and terminology

Named entities, which are often party terms, do not normally appear in dictionaries, and present difficulties for MT. Importation of the JRC-Names dataset[8] is on the task list. When completed, those names will be identified by Pre:D, along with their potential renderings in target languages that have been determined by JRC. Future work will incorporate additional named entities, as well as domain-specific terminology, from other data sources. Both names and terminology fall outside of the normal ambit of lexicography, but are part and parcel of any documents that might be subject to translation or NLP. It is therefore crucial that the terms can be discovered, their meanings determined, and that they be conveyed with their correct renderings in other languages.

2.5 User selections as learning data

With user permission, completed sentences will be stored as usage examples for their selected senses within the appropriate dictionary entries. This will result in a large corpus of human-disambiguated text in numerous languages, which can be used for

[8] https://ec.europa.eu/jrc/en/language-technologies/jrc-names

machine learning and other future NLP tasks. After future work integrates with computer assisted translation (CAT) output from registered translators, through a program such as Virtaal[9], the data can also be used as parallel text. Proposals for making use of this data are welcome.

3. Conclusion

Pre:D is a work in progress that discerns the meanings of terms within source documents, in order to identify appropriate vocabulary for translation purposes. The program relies on human-curated lexicons that link languages based on concepts that are known to overlap, rather than statistical best guesses. Furthermore, Pre:D asks users to delimit their units of analysis by selecting whether a word is an individual entity or part of a party term. Finally, the program tasks the user with selecting a term's meaning, rather than making a choice that has a high likelihood of being wrong. The amount of human involvement is much higher than the expectation of automaticity that has been inculcated by hands-free MT programs. Whether users will want to manage their translations to the level that Pre:D makes capable is an untested proposition. We foresee that some users will find the added precision invaluable, whereas others will prefer the instant but questionable results of current translation services. It may turn out that users combine the two approaches; for a quick translation that gives the gist of a document, truthy SMT that muddles concepts such as steaming hot and spicy hot might be good enough, whereas for mission-critical or professional work such as a restaurateur preparing a menu, the occasion will demand Pre:D. The ultimate goal is to merge the approaches within the software, so that users can home in on quality translations with both computer assistance and human control. With the Pre:D toolset overlying Kamusi's growing multilingual database, the groundwork is now in place for a system that can fill many of the gaps between the meanings intended within source documents and the words that are chosen for understanding those documents in many other languages.

4. References

Benjamin, M. (2016). Problems and Procedures to Make Wordnet Data (Retro)Fit for a Multilingual Dictionary. In Proceedings of the Eighth Global WordNet Conference, Bucharest, Romania: 27-33.

Benjamin, M. (2015). Crowdsourcing Microdata for Cost-Effective and Reliable Lexicography. In AsiaLex 2015 Conference Proceedings, Hong Kong: 213-221.

[9] http://virtaal.translatehouse.org/

Chen, W. (2016). The Discoursal Construction of the Lexicographer's Identity in a Learner's Dictionary: A Systemic Functional Perspective. International Journal of Lexicography (Advance Access) doi: 10.1093/ijl/ecw011

Fellbaum, C. (1998, ed.). WordNet: An Electronic Lexical Database. Cambridge, MA: MIT Press.

Hanks, P. (2015). Cognitive Semantics and the Lexicon. International Journal of Lexicography (2015) 28 (1): 86-106

Pustejovsky, J., and Rumshisky, A. (2008). Between Chaos and Structure: Interpreting Lexical Data Through a Theoretical Lens. International Journal of Lexicography 21 (3): 337-355.

Sag, I., Baldwin, T., Bond, F., Copestake, A. and Flickinger, D. (2002). Multiword Expressions: A Pain in the Neck for NLP . In Proceedings of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics (CICLing2002): 1-15

Tarp, S. (2008). Lexicography in the Borderland between Knowledge and Non-knowledge. General Lexicographical Theory with Particular Focus on Learner's Lexicography. Lexicographica Series Maior 134. Tübingen: Niemeyer.

Yong, H. and Peng, J. (2007). Bilingual Lexicography from a Communicative Perspective. TLRP 9. Amsterdam/Philadelphia: John Benjamins