|  |  |
| --- | --- |
| **Semantic Annotation Projection**   |  | | --- | | 050  051  052  053  054  055  056  057  058  059  060  061  062  063  064  065  066  067  068  069  070  071  072  073  074  075  076  077  078  079  080  081  082  083  084  085  086  087  088  089  090  091  092  093  094  095  096  097  098  099 |   **from English to Spanish and English to Dutch** |
|  |
| **LING 7800 Computational Lexical Semantics**  **Princess Dickens** |
|  |
|  |
|  |
|  |

Abstract

|  |
| --- |
| 000  001  002  003  004  005  006  007  008  009  010  011  012  013  014  015  016  017  018  019  020  021  022  023  024  025  026  027  028  029  030  031  032  033  034  035  036  037  038  039  040  041  042  043  044  045  046  047  048  049 |

This paper explains a semantic annotation projection project I carried out for my classes, LING 7800 - Computational Lexical Semantics and LING 6300 - Machine Learning, at the University of Colorado, Boulder during the Fall 2019 semester. Methods for pre-processing materials, assembling the projection pipeline, evaluating results, and conducting both linguistic and computational error analysis are described.

Introduction

With over one million words of English text manually annotated for predicate and argument structure, Propbank is a gold-standard resource for multiple Natural Language Processing applications, including tasks such as machine translation, question-answering, and information extraction. A similar database called VerbNet features the corresponding semantic roles of Propbank argument labels, essentially Who did What to Whom, How, When and Where? (Palmer et al., 2010; Palmer et al., 2005) While sufficient manual annotations exist for English, no such extensive corpus exists for other widely spoken languages, with the exception of German (Erk et al., 2003), Arabic (Zaghouani et al., 2010), Hindi (Vaidya et al., 2011), and a few others, which are much smaller in size. Because manual annotation requires a substantial amount of time and resources, computational linguists are faced with the dilemma of acquiring high-quality annotations for low-resources languages.

One solution that is currently being developed by computational linguists is Semantic Annotation Projection. Simply stated, annotation projection involves taking the gold standard argument structure annotations from sentences in a source language (usually English) and *projecting* them onto parallel sentences in another, “research poor” language. This way, corpora annotated for features such as part of speech, semantic role, dependency structure, named entities, etc. can be automatically generated in other languages and used in a variety of NLP applications.

CU Boulder Ph.D. student, Skatje Myers, is working on one such project for Russian. She has kindly assembled an English-Spanish projection pipeline and provided me with the results for the purpose of conducting linguistic error analysis. The Spanish parallel sentences were taken from the LORELEI Spanish language pack, provided by the DARPA program (Christianson, Duncan, & Onyshkevych. 2018). Skatje’s code essentially combines 3 separate input files to produce labels on the target language:

1. GIZA-format parallel sentences (English and Spanish) that have been word-aligned
2. Brat-format parsed sentences from Spanish that have been processed by the Universal Dependency Parser Pipeline
3. PropBank gold standard annotations for the English sentences

Figure : Image of Indo-European Family of Languages Chronological Flowchart, www.linguatics.com/indoeuropean\_languages.htm.

In addition to the English-Spanish projections, I used similar methods to test out some English-to-Dutch projections. The differences between the two projection pipelines were that the Dutch parallel sentences were obtained through a machine translator instead of crowdsourcing, a different automatic word aligner was aligner, and I wrote my own projection script for the English-Dutch projections.

|  |
| --- |
| 100  101  102  103  104  105  106  107  108  109  110  111  112  113  114  115  116  117  118  119  120  121  122  123  124  125  126  127  128  129  130  131  132  133  134  135  136  137  138  139  140  141  142  143  144  145  146  147  148  149 |

My project for this class was to evaluate the effectiveness of Skatje’s and my own projection scripts for Semantic Annotation Projection from English to Spanish and English to Dutch by identifying patterns of error. My job was then to try and explain why these errors occurred for certain semantic domains, parts of speech, types of clauses, etc., and consider ways to improve the projection model. Some researchers in the literature have tried to improve their models, for example, by aligning constituents instead of individual words, or have used a filter to remove problematic words from the sentences altogether (Pado & Lapata, 2009).

Predictions

In this task, I had two projections to evaluate, one from English to Spanish, and the other from English to Dutch. While English, Dutch, and Spanish all belong to the Indo-European language family, English and Dutch are closer in genealogy, that is, they belong to the same more recently localized branch, West Germanic, while Spanish does not, as it is Italic. Thus, it stands to reason that English and Dutch should have more similar grammatical structures and more vocabulary in common. This is significant since the process of annotation projection depends on the word aligner’s effectiveness in matching corresponding words in parallel sentences. If English and Dutch are more similar than English and Spanish, the word aligner should be able to identify these corresponding parts more accurately for Dutch and there should be fewer argument structure shifts, such as the nominalization of a verb (*The Routledge Handbook of Historical Linguistics* 2015). Thus, I predicted that the English-to-Dutch projections would have a *higher* accuracy than those from English-to-Spanish.

A screenshot of a cell phone

Description automatically generated

|  |
| --- |
| 150  151  152  153  154  155  156  157  158  159  160  161  162  163  164  165  166  167  168  169  170  171  172  173  174  175  176  177  178  179  180  181  182  183  184  185  186  187  188  189  190  191  192  193  194  195  196  197  198  199 |

Data

* 1. DARPA LORELEI

The parallel English and Spanish sentences used in this project were obtained through the Linguistic Data Consortium, hosted by UPenn, which manages LORELEI language packs created through DARPA, and is available to licensed members only. The Defense Advanced Research Projects Agency (DARPA) is a US Government agency for conducting research and providing aid and intervention around the world. According to (Christianson, Duncan, & Onyshkevych 2019), in the past 25 years, DARPA has deployed U.S. government workers in at least 250 humanitarian efforts, where over 800 languages are spoken. Through DARPA, the US Government funds a research program called Low Resource Languages for Emergent Incidents (LORELEI), whose aim is to not only reduce reliance on manually translated or manually annotated corpora, but also create tools to help workers quickly gain situational awareness from low-resource languages into English, such as names, locations, events, and sentiments. This involves using high-resource languages to gain information on related languages and ensuring that tools are fast and language-agnostic, goals the program has dubbed “Leveraging Language-Universal Resources” and “Projecting from Related-Language Resources” (Christianson, Duncan, & Onyshkevych 2019).

|  |
| --- |
| 200  201  202  203  204  205  206  207  208  209  210  211  212  213  214  215  216  217  218  219  220  221  222  223  224  225  226  227  228  229  230  231  232  233  234  235  236  237  238  239  240  241  242  243  244  245  246  247  248  249 |

* 1. Crowdsourced Spanish translations

Spanish is one of the LORELEI representative languages. Therefore, obtaining the English-Spanish parallel sentences for the English-to-Spanish pipeline, as Skatje reports, was as simple as downloading the LORELEI Spanish language pack from the LDC repository. Each LORELEI language pack includes at least one million words of text that has been translated from English into the target language and vice versa (Strassel & Tracey. 2016). These include news articles, a phrasebook of informal colloquialisms, and an elicitation corpus designed to prompt speakers to use a diverse set of linguistic structures. For the purposes of my investigation, I chose to use the first 220 parallel sentences from the English and Spanish Phrasebook**.** These weremanually produced native speaker translations crowdsourced by the U.S. Government.

* 1. DeepL Dutch translations

Because Dutch is not one of the LORELEI representative languages, I used a machine translator called DeepL to translate the same English sentences from the English-Spanish set to produce parallel English-Dutch sentences. DeepL, whose creators also own Linguee, uses neural networks and parallel corpora in their machine translation model (Himmelein 2019). According to (Ueli Reber 2019), the performance of DeepL is comparable to that of Google Translate in that it boasts a larger vocabulary and may even be more precise in translating whole documents.

1. Projection Pipeline

|  |
| --- |
| 250  251  252  253  254  255  256  257  258  259  260  261  262  263  264  265  266  267  268  269  270  271  272  273  274  275  276  277  278  279  280  281  282  283  284  285  286  287  288  289  290  291  292  293  294  295  296  297  298  299 |

* 1. EFMARAL for Spanish

The first step in the projection pipeline after obtaining parallel sentences is to align these sentences using an automatic word aligner. In order for this to work, sentences must first be pre-processed with tokenization, that is, separating all words and punctuation with a space, converting to lowercase, and formatting the sentences as in the following example:

source sentence one . ||| target sentence one .

source sentence two . ||| target sentence two .

For the English-to-Spanish projections, the automatic word aligner used was Efmaral, which according to Östling & Tiedemann (2016), uses a Bayesian model with Markov Chain Monte Carlo (MCMC) inference. This is a method for increasing accuracy through a larger number of iterations. While slower than other automatic word aligners, Efmaral claims performance that is comparable to that of Giza++ and higher than Fast Align (https://github.com/clab/fast\_align). It can be imported as a Python library. In the Efmaral README, the authors explain that that the higher performance of Efmaral over Fast Align may be due to the fact that Fast Align limits its training to five iterations, while Efmaral increases the number of iterations proportionally to the corpus size. In other words, the smaller the corpus, the more iterations.

* 1. Fast\_Align for Dutch

For the English-to-Spanish projections, the automatic word aligner used was Fast Align ([Dyer](http://www.cs.cmu.edu/~cdyer), [Chahuneau](http://victor.chahuneau.fr) & [Smith](http://www.cs.cmu.edu/~nasmith) 2013). Fast Align is similar to Efmaral in that it makes use of an unsupervised machine learning technique and relies on statistical comparisons of token occurrences in parallel data. While much faster than other models due to its five iterations, its creators admit that it is not as accurate. Nevertheless, they report that it is useful for “downstream translations systems on a variety of language pairs.” Like Efmaral, parallel sentences used as input must first be pre-processed with tokenization, conversion to lowercase, and separation with a triple pipe and white space. Skatje Meyers, who prepared and assembled every stage of the English-Spanish pipeline, used the methods for preprocessing outlined in (Ticconi 2012).

|  |
| --- |
| 300  301  302  303  304  305  306  307  308  309  310  311  312  313  314  315  316  317  318  319  320  321  322  323  324  325  326  327  328  329  330  331  332  333  334  335  336  337  338  339  340  341  342  343  344  345  346  347  348  349 |

* 1. UDPipe

For the purpose of visualization for the English-to-Spanish projections, an automatic parser called the Universal Dependency Pipe (UDPipe) was used. UDPipe is a trainable pipeline for tokenization, tagging, lemmatization and dependency parsing. UDPipe is language-agnostic and can be trained given annotated data in CoNLL-U format (See https://universaldependencies.org/format.html). Previously trained models can be found for a variety of languages, including Spanish. The UDPipe used for Spanish in this project was trained on the AnCora corpus, which is especially useful since it includes annotations for implicit arguments, (Taulé, Peris & Rodríguez 2016).

* 1. Spanish Pipeline

In summary, the process for creating projections from English to Spanish was the following:

1. UDPipe for part of speech tagging and dependency parsing
2. English-Spanish parallel phrasebook sentences from the LORELEI LDC Spanish language pack, with added PropBank predicate and argument annotations
3. EFMARAL word aligner for token indices
4. Projection script written by CU Boulder Ph.D. student, Skatje Myers.
5. Visualizations made by Skatje Meyers using the Brat annotation tool.

A screenshot of a cell phone

Description automatically generated

|  |
| --- |
| 350  351  352  353  354  355  356  357  358  359  360  361  362  363  364  365  366  367  368  369  370  371  372  373  374  375  376  377  378  379  380  381  382  383  384  385  386  387  388  389  390  391  392  393  394  395  396  397  398  399 |

* 1. Dutch Pipeline

1. English phrasebook sentences extracted from the parallel sentences in the LORELEI LDC Spanish language pack, with added PropBank predicate and argument annotations
2. Machine translations of the English phrasebook sentences using DeepL
3. Fast Align word aligner for token indices
4. My own projection script, with visualizations as an output

**A screenshot of a cell phone

Description automatically generated**

1. Evaluation
   1. Projection Accuracy

I considered a projection correct if the predicate and obligatory arguments were labeled correctly, except for a few special cases. In Spanish, for example, I did not mark a sentence wrong if there was no arg0 label when there was inflection for person on the verb. I only marked it wrong when the subject pronoun was present. I did, however, mark a sentence wrong if the direct or indirect object pronoun was attached to the verb but no arg1 or 2 was specified, since this information would be specified in the English argument annotations. Another special case is when the argument labels on the English sentences were incomplete. If only the predicate was labeled on the English sentence, then that was the only label I looked for on the target sentence, for both Spanish and Dutch. I evaluated 200 Spanish and 200 Dutch sentences manually. These were predominantly the same sentences for both languages, with up to ten different sentence pairs.

|  |
| --- |
| 400  401  402  403  404  405  406  407  408  409  410  411  412  413  414  415  416  417  418  419  420  421  422  423  424  425  426  427  428  429  430  431  432  433  434  435  436  437  438  439  440  441  442  443  444  445  446  447  448  449 |

I accepted some labels that could be considered incorrect on single words if they were part of a phrase. For example, in “eso suena muy divertido,” the attribute *divertido* is correctly labeled as the arg2 and *muy* serves as a modifier. Because both of them together form a phrase, I accepted the arg2 label on *muy* as well.

A picture containing screenshot

Description automatically generated

* 1. Pipeline Errors

The two main differences in the English-to-Spanish and English-to-Dutch pipelines were the translations and the automatic word aligners. Evaluating the word aligner for each source to target projection was more difficult than evaluating overall accuracy because of the format of the index mappings. A better visualization of these mappings would have made evaluation more straightforward and palatable without an automatic evaluation tool. My method for calculating the accuracy of the index mappings was to check for the number of incorrect mappings out of what would have been the total correct mappings, if one considers the number of correct mappings to be the number of tokens in the source sentence. I then checked for translation shifts that altered the argument structure, even if semantic information was preserved. I noted what likely led to these incorrect mappings, such as an arg1 in the source language becoming a predicate in the target language. Since overall accuracy was measured just by the correct projection of the predicate minimally and then occasionally the obligatory arguments, overall projection accuracy does not always correspond to word alignment accuracy.

1. Results

|  |
| --- |
| 450  451  452  453  454  455  456  457  458  459  460  461  462  463  464  465  466  467  468  469  470  471  472  473  474  475  476  477  478  479  480  481  482  483  484  485  486  487  488  489  490  491  492  493  494  495  496  497  498  499 |

* 1. Projection Accuracy – Spanish

Overall accuracy for Spanish is 46%. Apart from semantic role projection, significant translation issues were identified. Translation errors were found in 30% of all Spanish sentences. These translation errors were typical of native speakers, such as spelling mistakes, missing or superfluous accent marks, ungrammatical structures, and non-standard vocabulary typical of bilingual English and Spanish speakers. This is indicative of an annotation issue. Only 13.5% of sentences that were found to have a translation error had the correct semantic role labels. As for other issues, 37.5% of projections were due to projection issues only.

|  |  |  |
| --- | --- | --- |
| Spanish | Count | Figure |
| Number of Spanish sentences with translation errors | 60 | 0.3 |
| Number incorrect projections due to SRL only | 75 | 0.375 |
| Number correct projections despite translation issues | 27 | 0.135 |
| Total correct projections | 92 | 0.46 |
| Total entries | 200 |  |

* 1. Projection Accuracy – Dutch

Overall accuracy for Dutch is 42.5%. There were no translation issues that I could identify for the Dutch sentences.

|  |  |  |
| --- | --- | --- |
| Dutch | Count | Figure |
| Total correct projections | 85 | 0.425 |
| Total entries: | 200 |  |

* 1. Linguistic Patterns of Error – Spanish

|  |
| --- |
| 550  551  552  553  554  555  556  557  558  559  560  561  562  563  564  565  566  567  568  569  570  571  572  573  574  575  576  577  578  579  580  581  582  583  584  585  586  587  588  589  590  591  592  593  594  595  596  597  598  599 |

|  |
| --- |
| 500  501  502  503  504  505  506  507  508  509  510  511  512  513  514  515  516  517  518  519  520  521  522  523  524  525  526  527  528  529  530  531  532  533  534  535  536  537  538  539  540  541  542  543  544  545  546  547  548  549 |

In the English-to-Spanish projections, there were two sentence types for which projection was especially successful. The first type, “wh” questions, was unsurprising, since there is a rather small range of question words which could appear in the data, and these correspond well to their English counterparts.

A picture containing screenshot

Description automatically generated

The second sentence type, which involved modal auxiliaries with an infinitive main verb, was unexpectedly successful. The structure of this sentence type in Spanish is different from its parallel English type in that it involves morphological inflection for person on the verb, making the arg0 harder to identify. English, on the other hand, is more analytic, and has a separate subject pronoun for this function. Thus, the mapping from English to Spanish would be a many-to-one mapping in which the arg0 and modal auxiliary in English is mapped to just one word in Spanish. I expected this to be more difficult for the pipeline to handle, but most sentences of this type received correctly projected argument labels. My assumption is that the word aligner in the Spanish projection, Efmaral, was able to associate subject pronouns with certain conjugated verbs. Since there is a relatively limited number of modal auxiliaries in both English and Spanish, this does not seem like such a difficult task with representative data and enough training epochs.

A screenshot of a cell phone

Description automatically generated

As for shortcomings, there were certain errors that seemed to stem from some uniquely Spanish constructions. These included “se accidental” constructions and dative experiencer verbs. “Se accidental” is a kind of passive construction in Spanish in which inanimate objects take on an agentive role. This is often used to highlight the accidental nature of an event and possibly to remove blame. One example of this would be “Se te perdió algo” (*Did you lose something?)* Literally translated, this would be rendered “[Something] [lose itself] [on you]?” This is clearly a construction that would require more tools than a simple word aligner. The predicate *lose* in English takes an arg0 (you) and an arg1 (something) while in the parallel Spanish sentence, the roles are reversed. The predicate *perdió* is marked for an arg0, but this arg0 corresponds to the arg1 in English. It is also reflexive, making the arg1 *se* a referent to *something*. Later, a dative arg2 is added, which is the *te* in this sentence. Projected correctly, the mapping would be the following:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Did | you | lose | something | ? |
|  | Aux | Arg0 | Predicate | Arg1 | punct |
| ¿ | Se | te | perdió | algo | ? |
| punct | Arg1 | **Arg2** | Predicate | Arg0 | punct |

Another problematic construction was that of dative experiencer verbs. Like the *se accidental* construction, this involves a reversal of the arg0 and arg1 roles from English to Spanish and requires an additional arg2. One example of this would be “she didn’t seem to mind,” where *she* is the arg0. In the Spanish parallel sentence, this becomes “Parece que [a] ella no le importó (*It seems, it wasn’t important to her*), where she becomes an arg2 of the predicate, instead of the arg0. It should be noted that this is a bad translation, since the *personal a* is missing, which could indicate that there’s an indirect object present.

A picture containing screenshot

Description automatically generated

* 1. Linguistic Patterns of Error – Dutch

There are three constructions for which the pipeline performed exceptionally well for Dutch projections. These were 1) simple, verb-2nd active sentences, 2) polar questions, and 3) commands. It is clear that the reason these proved to be unproblematic was because of their almost identical structure to their parallel sentences in English in both number of sentence elements and word order. Indeed, the mapping for these sentences is almost always comprised of one-to-one alignments, as in the following examples:

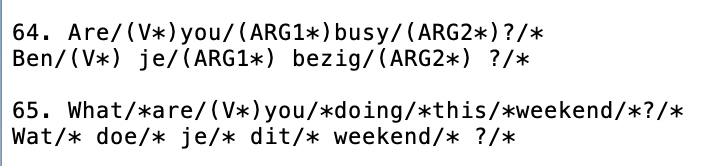
|  |
| --- |
| 600  601  602  603  604  605  606  607  608  609  610  611  612  613  614  615  616  617  618  619  620  621  622  623  624  625  626  627  628  629  630  631  632  633  634  635  636  637  638  639  640  641  642  643  644  645  646  647  648  649 |

**Simple, verb-2nd, active sentence**

A close up of a logo

Description automatically generated

**Polar question**



**Command**

A close up of a logo

Description automatically generated

Less successful were projections for sentences with structural differences, such as sentences with past-tense verbs. In English, the simple past is used to refer to events which have been completed and have no observable effect on the present. For past events which continue to affect the present, the present perfect construction is used. In Dutch, however, the simple past and present perfect are largely interchangeable, and there is a preference for the present perfect to be used even in cases where no present effect is observable(See:http://www.dutchgrammar.com/en/?n=Verbs.Re11). Thus, nearly all simple past verbs in the English source sentences corresponded to present perfect constructions in the Dutch target sentences. This not only adds the complication of a new auxiliary verb, but also a radical change in word order, since the main verb in the Dutch sentences would be moved to the end of the sentence instead of being in its usual verb-2nd slot. This was a complication for which the pipeline was not prepared.

A screenshot of a cell phone

Description automatically generated

|  |
| --- |
| 650  651  652  653  654  655  656  657  658  659  660  661  662  663  664  665  666  667  668  669  670  671  672  673  674  675  676  677  678  679  680  681  682  683  684  685  686  687  688  689  690  691  692  693  694  695  696  697  698  699 |

While it is understandable that the pipeline would not be able to handle major structural shifts from source to target language, there were also problems that highlighted the lack of robustness of the word aligner, Fast Align. Nearly all Dutch sentences that were parallel to English sentences containing the verb *to be* as a clitic were improperly labeled.

A close up of a logo

Description automatically generated

* 1. Crosslinguistic Patterns of Error

It is interesting to note that there were constructions that proved to be problematic in both pipelines. These were reflexives, punctuation, and negation. As for reflexives, it is obvious why the target sentences didn’t receive the correct obligatory argument labels in Spanish or Dutch, namely, because the source sentences in English were not reflexive and therefore lacked the arg1 label entirely. In order for projection to be successful for this type of construction, an extra piece would have to be added to the pipeline to check for added arguments.

A screenshot of a cell phone

Description automatically generated

A close up of a logo

Description automatically generated

Punctuation and negation were two common problems which I did not expect. For a human, these correspondences between periods, commas, or question marks in the source and target sentences seem obvious. Likewise, there are only a few negative words and particles in Indo-European languages, so their correspondence should also be evident. It is possible that the poor performance seen in this investigation is due to a lack of training data in which negative markers appeared.

Lastly, I would like to discuss two constructions which worked well in the Spanish projections but not in the Dutch ones and offer an explanation for why I think this occurred. The first construction was that of modal-auxiliaries and an infinitive main verb. These constructions are similar in all three of the languages involved in this investigation. The English- Spanish sentences differ in that the Spanish modal auxiliary is marked for person, while in English there is a separate subject pronoun and no marking on the verb. Word order is identical in that the modal auxiliary and main verb are adjacent with the modal auxiliary first. The English-Dutch sentences differ in that the Dutch modal auxiliary and main verb are not adjacent (the main verb goes to the end of the sentence), though morphology and subject pronouns correspond closely. These differences were minor in both cases and shouldn’t have interfered with semantic role projections at all. I believe that the Dutch projections suffered a drop in performance for these sentences because of the poor performance of the word aligner, Fast Align. These sentences were generally longer than those with simpler constructions, increasing the potential for erroneous mappings by the word aligner. More investigation is needed to tell if using Efmaral would resolve this issue.

The second construction which worked well for Spanish projections was that of “wh” questions, or questions which include a question word, such as “where”, “what”, or “when.” These generally require argument labels, which would correspond to their answers. For example, “what” elicits a response that indicate an agent, thus its label is arg0. These question words in the target languages are close to their English counterparts in both form and argument structure. Thus, the poor performance in the Dutch projections possibly indicates the need for more representative training data in the word aligner.

1. Discussion

Overall accuracy in both projection models is abysmal at 46% for Spanish and 43% for Dutch. While considerations into different elements of the pipeline need to be taken into account, it is clear that many errors are due to radical changes in argument structure from source to target language. In order to mitigate these errors, more comprehensive checks need to be added to each respective pipeline. Problems common to both pipelines, such as with the mislabeling of punctuation and negative markers highlight the need for more language-agnostic tools in argument identification. At the same time, problems unique to each target language demonstrate the complexity of handling major semantic difference even in closely related languages.

There are a few caveats involving data, methods, and evaluation, which should be taken into account when considering the validity of the results observed in this investigation.

First of all, translation errors were abundant in the Spanish language pack. Around 30% of the Spanish sentences I observed contained translation errors such as misspellings, absent or superfluous accent marks, and non-standard vocabulary or constructions. These mistakes are important to note since they are unlikely to occur multiple times in the training data and in the case of accents, can mean the difference between a determiner and a predicate.

Second, it should be noted that a different word aligner was used in each pipeline, making my results comparison less trustworthy. In order to evaluate the robustness of the methods described in this paper for multiple languages, the same tools should be used throughout the entire process.

Finally, it is important to recognize the limitations of manual evaluation. This created a major constraint on the number of examples, the number of arguments for each example, and the number of correct mapping for my alignment completeness metric I was able to assess. Moreover, manual evaluation can lead to instances of human error. Thus, a more uniform approach to implementing a semantic annotation projection pipeline with higher quality data and automatic evaluation would yield more precise results for a better understanding of the task and its limitations.

References

Aminian, M., Mohammad S. R., & Mona D. 2017. *Transferring Semantic Roles using Translation and Syntactic Information.*

Anders, Björkelund, Love, Hafdell, & Pierre, Nugues. 2009. *Proceedings of the Thirteenth Conference on Computational Natural Language Learning* (CoNLL 2009): Shared Task chapter Multilingual Semantic Role Labeling. As- sociation for Computational Linguistics.

Vaidya, Ashwini, Choi, Jinho, Palmer, Martha, & Bhuvana, Narasimhan. 2011. Analysis of the hindi proposition bank using dependency structure. In *Proceedings of the 5th Linguistic Annotation Workshop*, pages 21–29, Portland, Oregon, USA. Association for Computational Linguistics.

Brat rapid annotation tool. (Accessible at <https://brat.nlplab.org/introduction.html>)

Wang, Chenguang, Akbik, Alan, Chiticariu, Laura, Li,Yunyao, Xia, Fei & Xu, Anbang. 2017. Crowd-in-the-loop: A hybrid approach for annotating semantic roles. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1914–1923, Copenhagen, Denmark. Association for Computational Linguistics.

Dyer, Chris, Chahuneau, Victor & Smith, Noah A. 2013. A Simple, Fast, and Effective Reparameterization of IBM Model 2. In *Proc. of NAACL*.)

Christianson, Caitlin, Jason Duncan, & Boyan Onyshkevych. 2019. "Overview of the DARPA LORELEI Program." *Machine Translation* 32 (1): 3-9.

Georgi, Ryan, Fei Xia, and William D. Lewis. 2016. "Capturing Divergence in Dependency Trees to Improve Syntactic Projection."*Language Resources and Evaluation*, vol. 48, no. 4, 2014; 2016, pp. 709-739.

Hajič, Jan & Ciaramita, Massimiliano & Johansson, Richard & Kawahara, Daisuke & Màrquez, Lluís & Meyers, Adam & Nivre, Joakim & Padó, Sebastian & Štěpánek, Jan & Straňák, Pavel & Surdeanu, Mihai & Xue, Nianwen & Zhang, Yi. 2009. *The CoNLL-2009 shared task: Syntactic and semantic dependencies in multiple languages.* Proceedings of the Thirteenth Conference on Computational Natural Language Learning: Shared Task. 1-18.

Himmelein, Gerald. 2019. “DeepL: The New Gold Standard in Online Translation?” DeepL: The New Gold Standard in Online Translation?, SoftMaker, 3 June 2019, [www.softmaker.com/en/blog/bytes-and-beyond/deepl-the-new-gold-standard](http://www.softmaker.com/en/blog/bytes-and-beyond/deepl-the-new-gold-standard).

<http://clic.ub.edu/corpus/en>

<http://verbs.colorado.edu/semlink/>

Indo-European Family Of Languages Chronological Flowchart, [www.linguatics.com/indoeuropean\_languages.htm](http://www.linguatics.com/indoeuropean_languages.htm).

Erk, Katrin, Kowalski, Andrea, Padó, Sebastian & Pinkal, Manfred. 2003. Towards a resource for lexical semantics: a large German corpus with extensive semantic annotation. 2010, In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 537–544, Sapporo, Japan. Association for Computational Linguistics.

Linguistic Data Consortium (LDC) https://www.ldc.upenn.edu/about

Palmer, Martha, Gildea, Daniel & Xue, Nianwen. 2010. Semantic role labeling. *Synthesis Lectures on Human Language Technologies*, 3(1):1–103.

Palmer, Martha, Gildea, Daniel & Kingsbury, Paul. 2005. The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics, Volume 31, Number 1, March 2005*.

Östling, Robert & Jörg Tiedemann. 2016. "Efficient Word Alignment with Markov Chain Monte Carlo." *The Prague Bulletin of Mathematical Linguistics* 106 (1): 125-146.

Pado, S., & Lapata, M.. “Cross-Lingual Annotation Projection for Semantic Roles.” Journal of Artificial Intelligence Research 36 (2009): 307–340. Crossref. Web.

Strassel, Stephanie, & Jennifer Tracey. 2016. “LORELEI Language Packs: Data, Tools, and Resources for Technology Development in Low Resource Languages.” ACL Anthology, European Language Resources Association (ELRA). www.aclweb.org/anthology/L16-1521/.

Taulé, Mariona, Peris, Aina, and Rodríguez, Horacio. 2016. "Iarg-AnCora: Spanish Corpus Annotated with Implicit Arguments." Language Resources and Evaluation 50, no. 3: 549-584.

The Routledge Handbook of Historical Linguistics. 2015. Ringgold, Inc. pp. 645-656.

Ticconi, Fabio. 2012. *How to Do a Word Alignment with GIZA++ or MGIZA++ from Parallel Corpus.* 5 Mar. 2012, fabioticconi.wordpress.com/2011/01/17/how-to-do-a-word-alignment-with-giza-or-mgiza-from-parallel-corpus/.

Reber, Ueli. 2019. Overcoming Language Barriers: Assessing the Potential of Machine Translation and Topic Modeling for the Comparative Analysis of Multilingual Text Corpora, Communication Methods and Measures, 13:2, 102-125, DOI: 10.1080/19312458.2018.1555798

Vania, Clara, Yova Kementchedjhieva, Anders Søgaard, & Adam Lopez. 2019. "A Systematic Comparison of Methods for Low-Resource Dependency Parsing on Genuinely Low-Resource Languages."

Zaghouani, Wajdi, Diab, Mona, Mansouri, Aous, Pradhan, Sameer & Palmer, Martha. 2010. The revised arabic propbank. In *Proceedings of the Fourth Linguistic Annotation Workshop*, pages 222–226, Uppsala, Sweden. Association for Computational Linguistics.

Wijffels, Jan. 2019. UDPipe Natural Language Processing - Text Annotation. (Accessible at https://cran.r-project.org/web/packages/udpipe/vignettes/udpipe-annotation.html)

Zennaki, O., et al. 2019. “A Neural Approach for Inducing Multilingual Resources and Natural Language Processing Tools for Low-Resource Languages.” *Natural Language Engineering*, vol. 25, no. 1, 2019, pp. 43–67., doi:10.1017/S1351324918000293.