ChAnot: An Intelligent Annotation Tool for Indigenous and Highly Agglutinative Languages in Peru

Rodolfo Mercado-Gonzales, José Pereira-Noriega, Marco Sobrevilla, Arturo Oncevay
Research Group on Pattern Recognition and Applied Artificial Intelligence
Departamento de Ingeniería, Pontificia Universidad Católica del Perú, Lima, Peru
{rmercado, msobrevilla, arturo.oncevay}@pucp.edu.pe, jpereira@pucp.pe

Abstract

Linguistic corpus annotation is one of the most important phases for solving Natural Language Processing (NLP) tasks, as these methods are deeply involved with corpus-based techniques. However, meta-data annotation is a highly laborious manual task. A supportive alternative requires the use of computational tools. They are likely to simplify some of these operations, while can be adjusted appropriately to the needs of particular language features at the same time. Therefore, this paper presents ChAnot, a web-based annotation tool developed for Peruvian indigenous and highly agglutinative languages, where Shipibo-Konibo was the case study. This new tool is able to support a diverse set of linguistic annotation tasks, such as word segmentation, POS-tag markup, among others. Also, it includes a suggestion engine based on historic and machine learning models, and a set of statistics about previous annotations.

Keywords: Annotation Tool, Corpus Annotation, Peruvian Indigenous Languages, Shipibo-Konibo

1. Introduction

The research field of Natural Language Processing (NLP) aims the analysis automation, representation and generation of human language through computational techniques (Cambria and White, 2014). These methods usually are based on supervised machine learning approaches, which employ the analysis of corpus, composed of annotated samples, and other strategies for reaching a particular NLP goal. The samples within a corpus need to be enriched with additional information or meta-data, plus the correct solution regarding the NLP task, generating an annotated linguistic corpus (Müller and Strube, 2006).

Nevertheless, the development of an annotated linguistic corpus might be a highly time-consuming task. The process requires mainly human effort, which is performed by linguists or experts in a particular language. While the corpus is more specialized or the language is less widespread, the expertise requirements will increase. In that way, an easier specialized annotation tool is what is needed. (Müller and Strube, 2006).

This paper introduces ChAnot, a new web-based annotation tool, which is focused in Peruvian indigenous and highly agglutinative languages. ChAnot enables the corpus annotation for various NLP tasks, such as Morphological Analysis (lemma and affixes segmentation), Part-of-Speech tagging, Name Entity Recognition and Syntactic Analysis (using a BRAT[1] interface (Stenetorp et al., 2012)). The interactive interface and architecture are adjusted appropriately to the needs of the Peruvian indigenous languages and humans annotators. Moreover, ChAnot includes a suggestion engine based in machine learning models and a set of important statistics about historic annotation.

The text below is organized as follow. Section 2 presents some existing annotation tools. Then, the main and distinctive functionalities in ChAnot are detailed in Section 3. Later, Section 4 presents briefly features regarding the Peruvian indigenous languages, focusing in Shipibo-Konibo, which is the case study language. Finally, conclusions and future works are discussed.

2. Related Work

Due to the increasing of NLP applications during the last years, an important number of annotation tools have been developed. Each tool has features according to specific objectives. Therefore, there are annotation tools supporting a set of NLP tasks independently of the language, while other are focused in a specific kind of languages.

One of the mainly important annotation software is MMAX2[2], which is a GUI-based tool that, like most of the other tools, lets the users to select a portion of text and annotate some properties over it (text span annotation). This process enables the markup of POS tag, word senses, coreference, dependency relations, among others. Besides, the XML format is used to store the meta-data (Müller and Strube, 2006).

Most of the current annotation tools are web-based, language independent, use machine learning for managing suggestions and support a variety set of text annotation tasks. Within this range, BRAT is likely to be the most popular one, which additionally includes high-quality annotation visualization and is fully configurable (Stenetorp et al., 2012). Another popular tool is WebAnno[3], which additionally offers annotation project management (including management of users and roles) (Yimam et al., 2013).

Likewise, as it was mentioned before, there are some annotation tools developed according to features of a specific language, such as Fassieh (Attia et al., 2009). This GUI-based tool lets the user to perform morphological, POS-tag, phonetic and semantic annotations for Arabic texts.

As it may be observed, the described tools have many features in common. Generally speaking, they are very helpful for different annotation tasks. Nevertheless, a customized

[1] Available in: http://brat.nlplab.org/

[2] Available in: http://mmax2.net/

[3] Available in: https://webanno.github.io

tool for the features and reality of Peruvian indigenous languages is what is pursuit in this study. The main motivations and specific functionalities are described in the next section.

3. ChAnot Annotation Tool

The main reason for developing a new tool was the need to exploit morphological-rich languages, due that word segmentation together with morpheme meta-data annotation (for POS-tagging or disambiguation) is not possible in most of the other tools, as far as it was noticed in the previous part. Another motivation is the lack of experience of the human annotators (linguists and native speakers), who have never done this kind of work before, and even some of them are beginners in the use of software.

In this context, ChAnot[4] is an intelligent web-based annotation tool focused in Peruvian indigenous and agglutinative languages, which allows text processing through morphological (lemma and affixes), morpho-syntactic (POS tag), named entity and syntactic annotation (with an interface integrated with BRAT).

The name of the tool is composed from the terms *Chana*, the native name of a bird that represents knowledge in the Shipibo-Konibo culture, and the first part of *anotador*, which means annotator in Spanish. The latter is the main official language in Peru, and the reason why the interface has included Spanish terminology.

Likewise, ChAnot was implemented using a client-server architecture that can be accessible from any modern web browser, using a back-end implemented in Java. The annotated data is stored on a MySQL database for most tasks, except for the syntactic annotation which is saved in BRAT format.

3.1. ChAnot Workflow

ChAnot was designed to perform two main annotation phases. The first phase is a morphological and morpho-syntactic annotation. Then, using the previous information, and with the help of a BRAT interface, enables a syntactic annotation phase. NER task annotation is also available, although not integrated in the main work flow.

- **Input Phase**
  ChAnot receives a plain text file in encoding UTF-8 with a sentence per line. Each sentence has two parts (separated by a vertical bar): the own sentence in indigenous language followed by the translation of this sentence in Spanish (translation is not necessary), as it is shown in Figure 1.

```
Rámara pápa Iquitoain iki | Ahora papá está en Iquitos
Bakísha ea kái | Voy a ir mañana
```

Figure 1: Input format sample

- **Annotation Phase**
  After uploading the input file to ChAnot, the user must select an annotation task to perform. The unique restriction is asked for performing syntactic annotation, as it is required a previous morphological and POS-tag annotation.

- **Output Phase**
  Finally, the annotation is saved in a database. It can also be exported in a XML file structured by sentences, words and affixes. Figure 2 presents an output XML sample.

```
<corpus>
  <sentence id="0" text="Escuela tapo non matsóti iki ." translation="Barreremos el suelo.">
    <word id="0" lemma="escuela" posTag="Nombre" subPosTag="Contable" token="Escuela"/>
    <word id="1" lemma="tapo" lemmaPred="tapo" posTag="Nombre" subPosTag="Contable" token="tapo"/>
    <word id="2" lemma="noa" posTag="Pronombre" subPosTag="Personal" token="non">
      <affix order="1" text="n" type="Clit. nominal"/>
    </word>
    <word id="3" lemma="matsóti" posTag="Verbo" subPosTag="Transitivo" token="matsóti">
      <affix order="1" text="ti" type="Sufijo verbal"/>
    </word>
    <word id="4" lemma="iki" posTag="Verbo Auxiliar" subPosTag="" token="iki"/>
    <word id="5" lemma="." posTag="Puntuación" subPosTag="" token="."/>
  </sentence>
</corpus>
```

Figure 2: Output format sample

3.2. ChAnot Functional Features

Unlike other tools, ChAnot enables a complete morphological annotation (lemma plus affixes segmentation with annotation) and posses a communication interface with BRAT for dependency syntax annotation using the previous meta-data. The main features of ChAnot are detailed below.

- **Accessibility**
  Since the intended users of ChAnot are people who are not very familiar with computers or technology, it is a must to have a tool that can be accessible from anywhere and without the need to install any complicated software. In order to accomplish that, ChAnot is accessible from any modern web browser.

- **User Management**
  Each human annotator has a private account to work in ChAnot. This account let them to manage their annotation files through a menu (see Figure 3). There is no crossover of information and no possibility of getting corrupted data by external users either.

- **Statistics**
  ChAnot generates a set of important statistic per annotation file and per users, such as the current number of annotated sentences or words, and even the average time that was spent for the annotation task of each sentence. These statistics could be very useful as a complexity metric in the evaluation of the historically annotated sentences. Besides, it will be useful to evaluate the progress of human annotators for further analysis.

- **Interactive Interface**
  In order to ease annotation tasks for users, the interfaces were designed to be intuitive and for transmitting information trough different colors. In most of the ChAnot interfaces, green color means processed or completely annotated, yellow refers to work in progress, while red represents a not processed task or data. Figures 3 and 5 illustrates the color usage.

  Likewise, all text and messages in this tool are in Spanish, because it is the most spoken language in Peru.

[4]Source code and video demo available in: chana.inf.pucp.edu.pe/resources/chanot

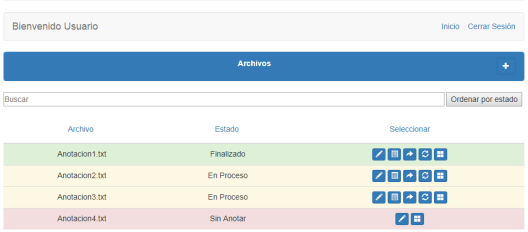

Figure 3: A user's main menu interface with a list of documents to annotate. The different colors represents the progress in each file.

- **Automatic Tokenization**
  ChAnot automatically tokenizes the input sentences in words, numbers, symbols and punctuation symbols. The split is performed for morphological, POS-tag and named entity annotation. This feature differs from most of other tools, which are mainly based on text span annotation.

- **BRAT Interface**
  Additionally, it was decided to take advantage of the BRAT capabilities in dependencies annotation for syntax, so a direct connection was configured from ChAnot.

  In the interface, each annotated sentence of the file is transformed into the BRAT files format. For that purpose, each word is split in its morphemes according the morphological annotation in ChAnot, which is entirely preserved (see the details in subsection 3.3). Figure 4 presents an annotation task in BRAT, where the POS-tags and word segmentation was obtained from the ChAnot interface.

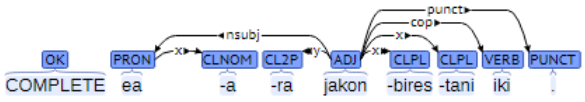

Figure 4: Dependencies annotation with a BRAT interface.

  It is important to highlight, that some dependency rules for annotation were introduced in the interface, in order to reduce the workload for the annotators in BRAT.

3.3. NLP Annotation Tasks in ChAnot

- **Morphological Annotation**
  In the morphological scope, annotation through ChAnot allows a complete word segmentation. For each word in a sentence, the lemma is introduced, and if applicable all the affixes could be added. The lemmas and morphemes must be annotated with their respective category, which are customizable regarding the language. Besides, the relative position of the sub-word units in the term is a requirement for finishing each word annotation. Figure 5 shows an example for a word with two affixes.

  As it is noted, the annotation includes the most relevant morphological information of each word plus the full meaning of each affix and their categories. The corpus can be exploited for other tasks such as word segmentation or morphological disambiguation.

- **POS-Tag Annotation**
  ChAnot allows Part-of-Speech tagging using a two-level predefined POS-tagset. The first level provides a general information about grammatical properties of a word, while the second describes more granulated information regarding the different kinds of the upper level.

  The tagset can be modified according the preferences or some specific language features. For the case study (Section 4), the definition of a POS-tagset aligned to the UD standard (Nivre et al., 2016) was an important factor.

- **Named Entity Annotation**
  As a complementary function, it is possible to identify if there is a named entity (NE) in the text. In ChAnot, the annotation is performed by introducing the specific NE category and a relative position tag, that discriminates single and multi-word entities in the sentence.

- **Syntactic Dependencies Annotation**
  A Universal Dependency (UD) Treebank (Nivre et al., 2016) is the main resource-like goal for the Peruvian indigenous languages (work in-progress). For that purpose, ChAnot includes a BRAT interface, as it was previously described.

3.4. Automatic Suggestions in ChAnot

ChAnot has integrated three different machine learning-based models that assist the workload of the annotators. One for the automatic identification of POS-tags, the second for the lemma prediction and a third for named entities recognition. The first two models are part of the ShipLemmaTagger toolkit (Pereira-Noriega et al., 2017), while the latter one is a newly hybrid model that uses a combination of rules and predictions based on previous annotations. All of these predictive models are implemented as Python web services in the server side. Furthermore, they are automatically updated in a periodic scheme with data coming from new annotations that are made by the users. As these models are frequently adjusted, the quality improvement of their suggestions is what is expected.
Furthermore, there are two kind of suggestions embedded in ChAnot: historic and machine learning-based. The former recalls the previous annotations (lemma, tags and affixes) that any user performed in the past, and presents the suggestion marking the word with a yellow fill. The latter one works as it was described before, and the suggestions are presented for entirely new words, highlighting them with a red fill. Figure 5 presents a sample.

4. Case Study: Shipibo-Konibo (shp)

4.1. Background

Peru presents a diverse culture map including many indigenous communities and cultures, who are minorities in the country. In order to support the preservation of their traditions and languages, the Ministry of Culture of Perú has identified 19 linguistic families including 47 indigenous languages, and 24 of them have been made official for government-service purposes. Among them, Shipibo-Konibo is the sixth language with the highest number of native speakers, with about 22 517 speakers, and is taught

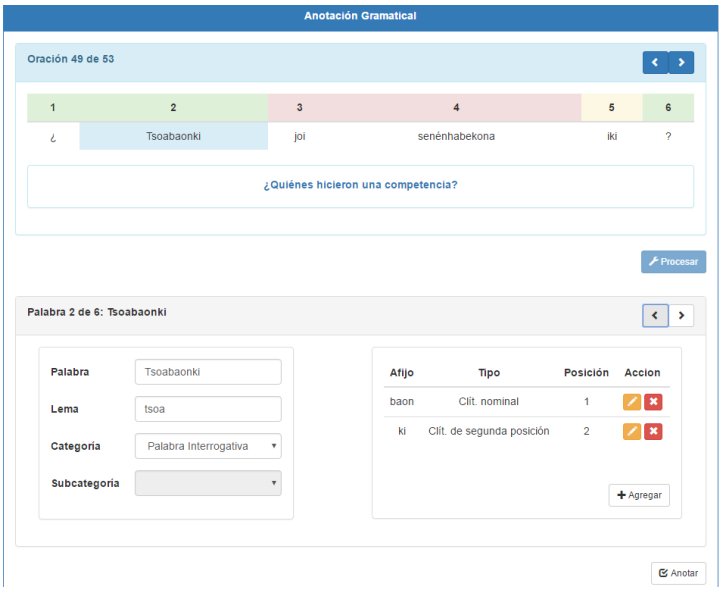

Figure 5: Morphological annotation sample (*Who made a competition?*): the word *Tsoabaonki* was split in *tsoa +baon +ki*, and each sub-word unit received additional information. Regarding the colors, the 3rd and 4th terms (red) are not annotated but they contain machine learning-based suggestions, while the 5th term (yellow) includes a historic-based suggestion.

in 299 public schools through bilingual educational programs (Ministerio de Cultura del Perú, 2016).

Shipibo-Konibo is a highly agglutinative language, with more than 100 suffixes and about 13 prefixes for word inflection (Valenzuela, 2003). Besides, there are not too many academic experts with experience in computational annotation tasks. In that context, ChAnot reflects as much information of this language for solving specific NLP tasks, while at the same time tries to be as easy as possible for unexperienced annotators.

4.2. Corpus Annotation and Evaluation

Using ChAnot, the experts could develop a corpus of 1630 annotated sentences, where each word within them contains: annotation of lemma, POS-tag, sub-POS-tag, and a list of all the affixes that compose the word. These affixes include category and relative positions. Besides, 204 and 78 sentences got named entity and dependencies (with the BRAT interface) annotations, respectively. However, the latter two tasks are not part of the analysis.

An evaluation of the machine learning-based suggestion engine was performed. It was simulated an increasing annotation scheme with automatically updated models in a 300-sentence block. In this sense, the accuracy achieved a peak of 74% for lemmatization, and 89% for POS-tagging (see Figure 6).

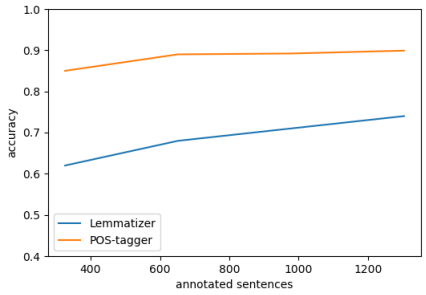

Figure 6: Evaluation of the lemmatizer and POS-tagger integrated in ChAnot

5. Conclusions and Future Work

The need for an annotation tool customized around the features of Peruvian indigenous languages allowed the design and implementation of ChAnot. The tool is likely to speed up the construction of linguistic corpora, by presenting suggestions based on past annotations samples, as well as predicting suggestions for new words with machine learning models that are frequently adjusted with newly annotated data. Furthermore, the wide range of functional features and annotations tasks are core points of the tool, which has the potential to scale easily for more complex jobs.

Future work is focused in the functional aspect. First, crowdsourcing, as it is included in WebAnno (Yimam et al., 2013), may be a relevant feature for enable the participation of more experts. However, the simultaneously integration of crowdsourcing and active learning schemes is what is really desired. The initial steps has been already made, as there are now some predictive models embedded in the tool. Another functional feature would be the inclusion of a spell-checker for the indigenous language, as it might work as a previous validation step for the required text to annotate (Alva and Oncevay, 2017).

Furthermore, there are new tasks that are expected to be annotated in the short term. Alignment is one of them, due to the presence of the translated text in Spanish (from parallel corpora). That could help enormously in the corpus-based machine translation experiments that has been performed recently (Galarreta et al., 2017). Finally, other NLP level tasks may be supported in the future, such as the semantic layer with word sense disambiguation annotations.

6. Bibliographical References

Alva, C. and Oncevay, A. (2017). Spell-checking based on syllabification and character-level graphs for a Peruvian agglutinative language. In *Proceedings of the First Workshop on Subword and Character Level Models in NLP*, pages 109–116.

Attia, M., Rashwan, M. A., and Al-Badrashiny, M. A. (2009). Fassieh, a semi-automatic visual interactive tool for morphological, PoS-Tags, phonetic, and semantic annotation of Arabic text corpora. *IEEE transactions on audio, speech, and language processing*, 17(5):916–925.

Cambria, E. and White, B. (2014). Jumping NLP curves: A review of natural language processing research. *IEEE Computational intelligence magazine*, 9(2):48–57.

Galarreta, A. P., Melgar, A., and Oncevay, A. (2017). Corpus creation and initial SMT experiments between Spanish and Shipibo-Konibo. In *RANLP (In-press)*.

Ministerio de Cultura del Perú. (2016). Base de datos de pueblos indígenas u originarios - Pueblos indígenas del Perú. http://bdpi.cultura.gob.pe/lista-de-pueblos-indigenas. Accessed: 2016-31-03.

Müller, C. and Strube, M. (2006). Multi-level annotation of linguistic data with MMAX2. *Corpus technology and language pedagogy: New resources, new tools, new methods*, 3:197–214.

Nivre, J., de Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajic, J., Manning, C. D., McDonald, R. T., Petrov, S., Pyysalo, S., Silveira, N., et al. (2016). Universal Dependencies v1: A multilingual treebank collection. In *LREC*.

Pereira-Noriega, J., Mercado-Gonzales, R., Melgar, A., Sobrevilla-Cabezudo, M., and Oncevay-Marcos, A. (2017). Ship-LemmaTagger: Building an NLP toolkit for a Peruvian native language. In *International Conference on Text, Speech, and Dialogue*, pages 473–481. Springer.

Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., and Tsujii, J. (2012). BRAT: A web-based tool for NLP-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '12, pages 102–107, Stroudsburg, PA, USA. Association for Computational Linguistics.

Valenzuela, P. (2003). *Transitivity in Shipibo-Konibo grammar*. Ph.D. thesis, University of Oregon.

Yimam, S. M., Gurevych, I., de Castilho, R. E., and Biemann, C. (2013). WebAnno: A flexible, web-based and visually supported system for distributed annotations. In *ACL (Conference System Demonstrations)*, pages 1–6.