Termolatío: Termolator for Spanish

Pauline Wee
New York University

Jhon Kim
New York University

Levith Andrade Cuellar
New York University

Abstract

The Termolator is an open-source terminology extraction system which so far works for English, Chinese, and French. This project seeks to address a significant gap in the Termolator's coverage by adapting its system to work for the Spanish language — one of the most spoken languages with almost a billion speakers worldwide.

The Spanish Termolator (*Termolatío*) works in five steps. First, *Termolatío* scrapes background and foreground corpora from Spanish Wikipedia with Beautiful Soup. It then processes the articles through part of speech (POS) tagging, noun chunking, and distributional ranking with an adapted system from previous versions of the Termolator. When this process is complete, *Termolatío* outputs a ranked term list of the most characteristic terminology of the foreground when compared to the background.

When evaluated by a Spanish speaker for precision, *Termolatío* resulted in 67% precision for terms more related to the foreground than the background category. However, further tests using a larger number of annotators and a wider test set need to be conducted in order to better benchmark *Termolatío's* performance compared to other automatic terminology extraction (ATE) systems.

1. **Introduction**

Automatic terminology extraction (ATE) systems are systems designed to extract domain-specific word sequences known as terminology from corpora on specialized topics. This terminology can then be used as keywords for information retrieval, domain-specific glossaries or ontologies, and summaries of large tests, making ATE systems extremely valuable for a range of applications in technology forecasting and document classification.

One of the most high-performing ATE systems available today is the Termolator. The Termolator is an open-source terminology extraction system that combines several novel and different ATE methods to get superior coverage and precision in terminology tagging. (Meyers et al, 2018).

Currently, the Termolator has working versions for English, Chinese, and French on the publicly available Github. A number of research papers have also tested Termolator's English version on various corpora. (Meyers et. al, 2015) (Pham, Pham, & Meyers, 2021) (Nordquist & Meyers, 2022) However, the Termolator does not yet cover Spanish, the official language of over twenty countries spoken by over one billion people. This project, entitled *Termolatío*, thus seeks to address this significant gap in the Termolator's coverage by adapting its system for the Spanish language.

In particular, *Termolatío* follows a similar process to the Chinese version of the Termolator. To obtain background and foreground documents, *Termolatío* uses a newly implemented scraper that cleans and prepares texts from Spanish Wikipedia using Selenium and Beautiful Soup. The articles then undergo part of speech (POS) tagging using a Python module called spaCy that provides Spanish coverage. Afterwards, the tagged terms are processed by a noun chunker adapted from the Chinese Termolator and a distributional ranking system extended to account for Spanish stopwords and POS tags. *Termolatío* then outputs a ranked term list of the most characteristic terms of the foreground when compared to the background.

To evaluate our system we enlist a native Spanish speaker to annotate the output of a test set for terminological relevance. We will then evaluate the results and provide recommendations for future research.

2. **Previous Work**

2.1 Automatic Terminology Extraction Systems (ATE)

While there are many definitions for "terminology" in the field of natural language processing (Rigouts Terryn et al., 2020), in this study, we refer to the definition used by Meyers et al, 2018, where terminology is distinguished as a word or phrase that a typical naive adult outside of the domain of the specialized field would not be expected to know the meaning of the term (Meyers et al., 2018). As units of knowledge in a specific field of expertise, these extracted terms can support and improve complex downstream tasks, e.g., information retrieval, machine translation, topic detection, and sentiment analysis. (Tran et al., 2023)

However, terminology is very labor intensive to extract manually, especially from large amounts of corpora. Thus, many Automatic Terminology Extraction Systems (ATE) have been designed to ease the effort of manually identifying terms from domain-specific corpora.

ATE systems can be trained on monolingual or multilingual datasets, and they can have varying features depending on their intended use. They can also be categorized into traditional non deep learning based systems and more recent deep learning based systems for terminology extraction. (Tran et al., 2023)

Notable traditional systems that do not employ deep learning include Termolator, the project upon which *Termolatío* is adapted, and Termostat (Drouin, 2003), a terminology extraction system that has a distributional component similar to Termolator's, but uses a fixed corpus as its background corpus for all foreground corpora (Meyers et. al, 2018). These systems identify terminology by evaluating linguistic and statistical features and combining multiple forms of information.

In more recent years, more novel approaches have employed deep-learning based and Transformer-based neural models to perform terminology extraction. These systems either use supervised or unsupervised methods and generally follow three steps: preprocessing, feature engineering, and term extracting classifier. (Tran et al., 2023)

While some research has found that traditional models outperform neural network classifiers on certain metrics for legal corpora (Howe et al., 2019), more recent surveys have found that neural models generally outperform machine learning models based on feature engineering by a large margin. However, these systems struggle to capture multi-word and nested terms, demonstrating a need to still improve these systems' training or design to handle such edge cases. (Tran et al., 2023)

Lastly, there are several methods to evaluate the output of ATE systems. These include intrinsic methods, which evaluate some property of the extracted list, usually through evaluation by a human expert in the specialized field, and extrinsic methods, which assess the quality of the extracting system by measuring performance improvement of another system or application that takes a terminology list as input. (Tran et al., 2023)

This project will use the intrinsic evaluation method used by Meyers et al. (2018), which involves randomly selecting 20 terms from each 20% interval in the ranked output. This selected list of 100 terms is then evaluated by a relevant domain specialist.

2.2 The Termolator

The Termolator is an open-source automatic terminology extraction system (ATE) first released in 2019 by Adam Meyers, Yifan He, Zachary Glass and Shasha Liao. (Meyers et al., 2018). The Termolator works by cleaning and scraping a set of documents separated into a more general “background” and a more specific “foreground” and identifying potential instances of terminology using a chunking procedure, similar to noun group chunking, but favoring chunks that contain out-of-vocabulary words, nominalizations, technical adjectives, and other specialized word classes. Its distributional component then ranks such term chunks according to several metrics including frequency in a more specific foreground than a more general background, a well-formedness score based on linguistic features, and a relevance score based on term appearance in articles and patents. Finally, it outputs a ranked list of most relevant terms, notably with an emphasis on words that are more specialized with regards to the foreground rather than the background. (Meyers et. al, 2018)

With these features, the Termolator has achieved upwards of 70% precision in identifying terminology for corpora in domains such as patents or biological research. (Meyers et. al, 2015) (Pham, Pham, & Meyers, 2021) (Nordquist & Meyers, 2022) However, these evaluations are mostly based on Termolator’s original English system, and thus do not fully reflect its precision for other languages, which include Chinese and French.

To create a Spanish version of Termolator, we originally planned to adapt the original English version of Termolator. However, the English Termolator had extensions that utilized the terms extracted from the base model. Since we hoped to create a base model for the Spanish Termolator, we decided to adapt the Chinese version as it did not have any base model-dependent developments aside from the accessor variety filter, which we did not have to implement or adapt for Spanish due to linguistic differences wherein Spanish words have clearly marked boundaries between words. We detail the steps of that program and how we adapted them to Spanish in the “Our System” section below.

3. Data Preparation

Our data was collected from Spanish Wikipedia with a scraper utilizing Selenium, a Python module for browser automation, and Beautiful Soup, a Python module that performs HTML parsing. We used the scraper to obtain two different sets of corpuses for development and testing. Since the distributional ranking component of *Termolatío* has been inherited from and trained accordingly for the task of terminology extraction by previous versions of the Termolator, we decided to focus on developing appropriate features for the Spanish version by using a development set and then testing the quality of our output by using a test set.

Our development set consisted of a background set on the topic of *chorizo*, a type of pork sausage enjoyed across Spanish speaking countries, and a foreground set on the topic of *platos con chorizo*, essentially dishes that contain *chorizo* as an ingredient. We opted for this development set given the wealth of cultural, Spanish-specific terminology that surrounds *chorizo* and the dishes that contain it.

Our test set consisted of a background on the topic of *tecnología*, technology, and a foreground on the subject of *tecnología médica,* medical technology. We opted for this combination of background and foreground to evaluate how our system would perform on the most common task it would be used for: discerning characteristic terms from a highly technical foreground and background combination.

3.1. Task Description

This project aims to create a Spanish version of the Termolator tool. *Termolatío*'s task, like that of previous versions of the Termolator, is to rank the most characteristic terms found on a foreground set of documents from those that can be found on a related but much wider set of background documents.

3.2. Our System

Termolatío approaches the task of term ranking between a foreground and background in four different steps mimicking the Chinese version of the Termolator.

First, *Termolatío* begins by obtaining and cleaning background and foreground sets from Wikipedia by using a custom web scraper that leverages Selenium and the Beautiful Soup library. Wikipedia was chosen as the corpora source because as an open-source intellectual resource, the ability to tag and rank terminology from its pages enables a broad category of people to efficiently navigate its heavily nested categories and hyperlinked pages.

Subsequently, *Termolatío* conducts Part-of-Speech (POS) tagging on the collected sets by using the spaCy library, a free open-source library for Natural Language Processing in Python. It features NER, POS tagging, dependency parsing, word vectors and more. spaCy was chosen for its high accuracy at POS tagging, which is 97% for English.

After this, *Termolatío* proceeds to conduct noun chunking on the texts by using an adapted version of the Chinese Termolator's noun chunker. This noun chunker explicitly favors chunks containing OOV and technical words compared to other less performant systems like Termostat. (Drouin, 2003). *Termolatío*'s noun chunker adapts this for Spanish so that it is capable of recognizing and utilizing Spanish POS tags for this task.

In the last stage, *Termolatío* uses a standard distributional ranking approach borrowed from previous versions of the Termolator. This process was adapted to utilize Spanish stop words and spacy POS tags to optimize relevance. A subsequent step in other versions of the Termolator, the Accessor Variety Filter, is omitted in the *Termolatío* system given that Spanish has distinct noun boundary characteristics and noun groups.

When finished, *Termolatío* outputs a refined term list, denoted by the ".out_term_list" extension, which is a file containing a list of terms enriched with weighted scores, providing context and a ranking of the most characteristic terms found in the foreground.

3.3. Annotation and Evaluation

To evaluate *Termolatío* we follow the same process that is utilized for evaluating the English version of the Termolator. To evaluate *Termolatío* we began by running the system with our test set of background documents on the subject of *tecnología*, technology, and our test set of foreground documents on the subject of *tecnología médica,* medical technology.

After obtaining the output — the weighted and ranked list of most characteristic terms found in the foreground set — we employed the intrinsic evaluation method used by Meyers et al. (2018). We created a Python script to randomly select 20 terms from each 20% interval in the ranked output, i.e. 20 from the top 20%, 20 from the 21st to 40th percentile and so forth. This selected list of 100 terms was then evaluated by a Spanish speaker who evaluated the terms with "yes" or "no" annotations. The Spanish speaker was instructed to answer "yes" whenever they deemed the term to be well-formed and characteristic of the foreground topic compared to the background topic.

4. **Results**

4.1. Scoring Metric

Precision was calculated by dividing the number of “yes” annotations provided by the Spanish speaker by 100, the selected number of representative terms from the output.

4.2. Test Set Results

Termolatío outputted a list of ~1500 terms when running our *tecnología* - *tecnología médica* background - foreground set. From those ~1500 terms, 100 terms were randomly selected and provided to a Spanish speaker for annotation. The annotated list was then scored based on the metric detailed in 4.1, resulting in a final precision of ~67%.

Below is a small sample of the annotated results. The “S” column contains the system output and the “A” column contains the annotator's decision.

Fig. 1: Sample of Annotation of Terminology Extracted from the *tecnología* - *tecnología médica* background - foreground set

Terminology Extracted	**S**	**A**
inflamatorias autoinmunes	YES	YES
alta abundancia	YES	YES
nueva ovulación	YES	YES
técnicas reproducción	YES	YES
segunda ortopedia	YES	YES
nueva narrativa	YES	NO
buen historial	YES	YES
gran numéro	YES	NO
últimos años	YES	NO
mayor calidad	YES	NO

5. **Discussion**

Termolatío’s precision for the *tecnología* - *tecnología médica* background - foreground set was scored at 67%. This metric was determined when comparing the system’s results with those of the annotator. It is important to note that the process of using an annotator is highly subjective; different annotators may have varying judgements regarding the relevance of the output terms to the subject of the foreground.

To ensure consistency, our annotator formed rules and assumptions about which terms should be considered correct or sufficiently characteristic of the foreground. The output terms were diverse, and were not all highly technical.

A large portion of the terms were evaluative in nature. They were related to probability, sample size, or referenced the outcomes of medical test results. Since these terms would typically be used in the field of medicine, and thus in the field of medical technology, the annotator counted these as correct. Examples include:

- *alta abundancia*
- *mayor probabilidad*
- *peores resultados*

The annotator, however, did not count all evaluative terms as correct. The annotator discarded terms that are widely used in other fields of science and/or technology. Particularly terms related to quality or size — they are simply not as characteristic to medicine or medical technology as the aforementioned terms are. Examples include:

- *gran complejidad*
- *grandes efectos*
- *igual tamaño*

The more straightforward, highly technical terminology were immediately counted by the annotator. These were primarily medical:

- *inflamatorias autoinmunes*
- *nueva ovulación*
- *segunda ortopedia*

Using an annotator to evaluate our results shed light on the difficulties of determining precision. It became evident that any annotator must form their own rules and assumptions to provide consistent results, a bias which is difficult to account for in analyzing our results. In order to improve the quality of our results and evaluation, using several annotators and standardizing some rules and assumptions would be the most effective way to improve results moving forward.

6. Error Analysis

As briefly mentioned in the results section, the biggest caveat to precision scoring was subjectivity. It is entirely up to the annotator to decide whether the system output was relevant to the passed corpora. Hence, the precision scores are entirely dependent on the annotator and therefore subject to human error. In order to counteract this limitation, we believe there are two main elements we can implement in further trials. First, having a larger number of Spanish annotators calculate precision scores and averaging it will help increase the reliability of precision. Secondly, having multiple trials with different foreground and background combinations will further help enhance the accuracy of the precision score.

In our algorithmic analysis of *Termolatío*, we scrutinized each component of the system for its efficacy in processing Spanish terminology. The Part-of-Speech (POS) tagger, adapted from spaCy, included a lower level of categorization but still maintained a high level of accuracy in identifying grammatical structures, which is crucial for effective noun chunking. However, we observed that the adapted noun chunker, while proficient in identifying standard Spanish noun phrases, occasionally struggled with compound and technical terms, leading to missed terminology. These insights suggest that while the foundational components are robust, there is room for refinement, particularly in enhancing the noun chunker's ability to handle complex terms in Spanish specifically.

Lastly, in our comparative analysis, we evaluated *Termolatío's* performance across different text types within our development set. When processing technical texts, *particularly* those with dense scientific jargon, Termolatío demonstrated higher precision in term extraction, likely due to the distinct and repetitive nature of technical terminology. However, in more general texts and topics, the precision tended to be slightly lower. This variance can be attributed to the broader linguistic diversity and subtler usage of specialized terms in these texts in addition to the size of each corpora. The system's current configuration, while adept at handling highly specialized language, requires further tuning to effectively process a wider range of text types with the same level of accuracy.

7. Conclusion

Termolatío demonstrates significant initial strides in adapting the Termolator for the Spanish language, a crucial step towards addressing a notable gap in terminology extraction resources for one of the world's major languages. Its aim was to explore the feasibility of such an adaptation and lay the groundwork for future enhancements in this area. While our results indicate a promising start, with *Termolatío* achieving a basic level of precision in extracting Spanish terminology, there is a need to further refine and develop the system.

Our system currently shows capability in handling technical texts but the variations in precision across different text types suggest that further tuning is necessary to improve its adaptability and accuracy. The reliance on human annotators for precision scoring also poses a challenge in achieving objectivity in evaluating the terminology-extraction outputs. This points towards a potential for a more systematic enhancement in scoring via enlarging our test cases, trials, and number of annotators. The error analysis conducted has shed light on specific areas for improvement within *Termolatío's* algorithmic framework. For example, adjusting components such as the Spanish noun chunker would be one of the key steps towards increasing the system's overall effectiveness.

8. Future Work

In the future, we plan to expand *Termolatío* with more features that would enable it to more effectively perform terminology extraction technology for Spanish corpora.

First, a more intuitive and comprehensive user interface would allow more users to easily input links or documents for processing as foreground or background. We would like to develop the current command line interface or make a graphic user interface accessible over the internet or elsewhere in order to make *Termolatío* easier to use for global audiences. Integrating the web scraping program into *Termolatío* such that the user is able to choose their foreground or background files is another goal that we wish to achieve with further development. In doing so, we also hope to make the web scraping and/or document upload process faster, easier, and more native, so that it would enable more corpora to be processed through *Termolatío.*

Another improvement would be to improve the individual components of *Termolatío* such as the noun chunker, POS tagger, and distributional ranking. For example, more robust features could be integrated to source and clean corpus texts, and the spaCy component's tagging accuracy could be improved by training it with hand-labelled POS tags for Spanish, especially given that Spanish might be slightly more out-of-domain than the far larger library of English corpora that spaCy was originally trained on.

A more complex system adapting more of the features of English Termolator would also likely help improve the scores of precision. We could then compare different versions and combinations of those components to find which have the best results.

Lastly, *Termolatío* could be applied to more sets of corpora in order to more effectively evaluate the system's precision and accuracy for diverse forms of text. For example, *Termolatío* could be applied to academic papers in Spanish, newspaper documents from large publications in Latin America, and Spanish literature in order to evaluate how it performs across these different forms of corpora. It could also be compared to corpora that have been previously annotated to extract terminology in order to benchmark its effectiveness compared to human evaluators.

References

Drouin, P. (2003). Term extraction using non-technical corpora as a point of leverage. Terminology, 9, 99–115. https://doi.org/10.1075/term.9.1.06dro

Drouin, P., Rigouts Terryn, A., Hoste, V., & Lefever, E. (2020). TermEval 2020: Shared task on automatic term extraction using the annotated corpora for term extraction research (ACTER) dataset. In Proceedings of the 6th International Workshop on Computational Terminology (pp. 85–94). Marseille, France: European Language Resources Association.

Howe, J. S. T., Khang, L. H., & Chai, I. E. (2019). Legal area classification: A comparative study of text classifiers on Singapore Supreme Court judgments. arXiv preprint arXiv:1904.06470.

Meyers, A., He, Y., Glass, Z., & Babko-Malaya, O. (2015). The Termolator: Terminology Recognition based on Chunking, Statistical and Search-based Scores. Workshop on Mining Scientific Papers: Computational Linguistics and Bibliometrics. Retrieved from http://ceur-ws.org/Vol-1384/paper5.pdf

Meyers, A., Pham, N., & Pham, L. (2021). Legal Terminology Extraction with the Termolator. Natural Legal Language Processing 2021. Retrieved from https://aclanthology.org/2021.nllp-1.16.pdf

Nordquist, S., & Meyers, A. (2022). On Breadth Alone: Improving the Precision of Terminology Extraction Systems on Patent Corpora. Natural Legal Language Processing 2022. Retrieved from https://aclanthology.org/2022.nllp-1.1.pdf

Tran, H. T. H., Martinc, M., Caporusso, J., Doucet, A., & Pollak, S. (2023). The Recent Advances in Automatic Term Extraction: A survey. arXiv preprint arXiv:2301.06767.