Enriching Natural Language Processing Systems with Semantic-Pragmatic Information through Communicative Intentions

María Miró Maestre[1]

[1]Department of Software and Computing Systems, University of Alicante, 03690 Alicante, Spain

Abstract

Communicative intentions are one of the linguistic elements that usually determine the content of any message we want to express. However, regardless of the high precision Natural Language Processing (NLP) systems are acquiring these days, thanks to the revolution derived from the explosion of the latest large language models (LLMs), these architectures still show a lack of appropriate training in order to detect the intention of a message correctly. For the purpose of improving these systems, the present research project aims to create a communicative intention annotation scheme based on the taxonomy presented in the Speech Act Theory. Such resource could help NLP architectures to consider communicative intentions as a starting point to classify any message depending first on the intention it reflects. With this aim, the scheme will be created with the help of an already annotated corpus in Spanish. Subsequently, we will test the scheme within a classification system so that we can verify the accuracy of the intention patterns detected. In this way, it will be possible to check if NLP systems are capable of identifying Spanish communicative intentions or even generate messages that reflect a given intention, therefore enriching the linguistic information these architectures can infer automatically.

Keywords

communicative intentions, speech acts, natural language processing, annotation scheme, classification system

1. Introduction and Motivation

Natural Language Processing (NLP) systems, and more concretely Natural Language Generation systems (NLG), are nowadays at their peak due to the evolution that the large language models have shown these last years. Therefore, we currently have at our disposal more and more precise classification and generation systems when it comes to identifying the linguistic patterns demanded in the text to be processed or generated. This is the case of the task of abstractive summary generation in the NLG research branch, or the detection of offensive or humorous messages if we focus on current NLP tasks. Both represent a few examples of how automatic learning systems are starting to correctly detect and generate more concrete and ambiguous linguistic features each time.

Nevertheless, despite the excellent results that these systems show when detecting linguistic patterns belonging to levels of analysis such as morphology, syntax, and semantics, there is still

Doctoral Symposium on Natural Language Processing from the Proyecto ILENIA, 28 September 2023, Jaén, Spain.

maria.miro@ua.es (M. M. Maestre)

0000-0001-7996-4440 (M. M. Maestre)

a long way to refine such architectures so that they can finally capture the natural aspect of a message. One of the elements that mainly helps to define the structure and, generally, the sense of a message is its communicative intention. In fact, two of the main scopes of research that are receiving a lot of attention nowadays in the context of Artificial Intelligence are the tasks of intention detection in the scope of NLP, and conscious text generation with the inclusion of external knowledge in the NLG branch of research.

These tasks are usually oriented normally to creating conversational agents where a human tries to keep a conversation with a robot. Consequently, researchers need to study the linguistic parameters that denote a given communicative intention and the best approach to integrate them into automatically generated responses so that they are created according to the contextual and linguistic requirements of the conversation. However, regardless of the diverse works we also find on intention detection in written textual genres, the scientific output is still reserved for a small percentage of languages, with English at the head.

Consequently, our goal with this research proposal is to create a communicative intention annotation scheme in Spanish to incorporate more semantic-pragmatic information in a NLG system thanks to the linguistic indicators of each intention that we have gathered in our guidelines for the annotation task. Moreover, this initial corpus will serve us as a base for testing several semi-supervised classification techniques to augment the final weight of the corpus. In this way, we will provide the NLP research community with a valuable linguistic resource for identifying more linguistic patterns automatically in a language other than English.

The remainder of this article is organised as follows: Section 2 focuses on the different approaches made in NLP in order to tackle the automatic classification of semantic-pragmatic elements of language, then Section 3 shows the main hypotheses and objectives planned for this research. Subsequently, we explain the methodology proposed for fulfilling each project task in Section 4, and Section 5 sets out the different research issues we may need to face throughout the experimentation. Finally, the bibliography used for this study is included at the end of the paper.

2. Related Work

Regardless of the numerous innovative techniques available nowadays to do research on different linguistic tasks within the scope of NLP, there are well-known traditional linguistic theories that still shed light on how to approach some of the most difficult NLP tasks to resolve. This is the case of the Speech Act Theory (SAT) founded by Austin [1] and extended by Searle [2, 3], who defended that language can serve as a means to perform actions depending on the uttered message. To verify so, it was Austin [1] who first investigated verbs to identify how they could denote actions on their own (called *performative verbs* or either describe reality (*descriptive verbs*). Subsequent to this first pragmatic division, Austin focused his research on one of the aspects that comprises the act of uttering a message: the illocutionary act (i.e., the intention of a utterance). With this element, he created a 5-fold typology of intentions, although it was Searle's Searle [2] modified version the one generally accepted by the research community, given its more thorough and well delimited approach (see Table 1).

Later on, Searle [3] also made a distinction between the types of intentions aforementioned,

Intention	Description	Examples
Assertives	we commit to the veracity of the message expressed	declare, manifest, conclude, explain, etc.
Directives	the speaker uses this type to make the listener do something	ask for, dare, invite, command, challenge, etc.
Commissives	they commit the speaker to do an action in the future	swear, promise, commit, intend, etc.
Expressives	they express the psychological state of the speaker with respect to a topic specified in the message	thank, forgive, excuse, congratulate, etc.
Declaratives	when uttering them we get the content of the message to coincide with reality, that is, the action is performed, or in Searle's own words: 'saying makes it so'	declare, designate, resign, marry, etc.

Table 1
Taxonomy of intentions according to the SAT.

known as *direct speech acts* because the relation between the meaning and the intention of the message is straightforward, and other type of illocutionary acts called *indirect speech acts*. In the latter, the relation between the message and the intention requires other inferential processes (i.e., cultural references, social context, etc.) to successfully interpret the intention of the message, as in those texts containing irony, sarcasm or rhetorical questions, among others.

Despite the difficulties that the inclusion of pragmatic elements inside NLP and NLG systems entailed, several studies focused on this linguistic level to make progress in these domains of computational linguistics [4, 5, 6]. Therefore, we now find manifold research works enriching systems with pragmatic knowledge, and more concretely with communicative intentions classifications, to improve their efficiency.

A very prolific area of research is that devoted to the study of computer-mediated communication (CMC) [7], which includes the study of the language used in different social media platforms in its research scope. Specifically, within the scope of social media users' communicative intentions, several works have been published these last years where the Speech Act Theory serves as a base to identify users' intentions in their tweets, as in Saha et al. [8] and Zhang et al. [9]. Even some researchers have proposed to mix several NLP tasks in CMC corpora as in [10, 11] or [12], where authors test the implication of both sentiment analysis and emotion recognition tasks when trying to detect the intention of a tweet. If we broaden our scope of research to further NLP tasks, the SAT taxonomy meant a key point for studying the best approach to develop systems that could also automatically identify text intentions in task-oriented conversational systems [13].

Nevertheless, it is clear that most research on this subject handles English documents so far. However, we can still find a few examples of works where Spanish speech acts are used to either improve task-oriented dialogues as in Martínez-Hinarejos et al. [14] and Caballero et al. [15], analyse pathological language extracted from clinical oral data in Spanish, as shown in Gallardo Paúls and Fernández Urquiza [16], or even study CMC travel blogs [17] and the types of speech acts found in the social network Facebook [18].

3. Main hypotheses and objectives

Because of the numerous NLP scenarios in which we can make use of the SAT nowadays, our research proposal is based on the creation of a communicative intention annotation scheme in Spanish to use it as a resource to solve some of the most currently studied tasks in our field. In this way, we will try to improve the ability of current language models when identifying more semantic-pragmatic aspects of language to successfully reproduce them. More concretely, the main research questions that support this project are:

- RQ 1) Which linguistic features can help us detect the intention of a given message in the Spanish language?
- RQ 2) Is it possible to identify those linguistic features in a CMC corpus, given its colloquial style and lexical variety?
- RQ 3) Can a language model learn to differentiate between different types of intentions with a training dataset, regardless of the ambiguities inherent to language?
- RQ 4) Can both sentiment analysis and emotion detection help to identify the communicative intention of a message with better precision?
- RQ 5) Does the automatic annotation of communicative intentions benefit NLP applications such as an automatic text generator?

4. Methodology and proposed experiments

To integrate communicative intentions in some NLP current tasks to enrich systems with further semantic-pragmatic information, we will focus on Searle's classification of *direct speech acts* as explained in Section 2 and other linguistic features that also reflect the intention of the message straightforwardly. To create the corresponding annotation scheme, and to test its validity in an NLP application, several linguistic resources and computing tools were used to complete each of the experimentations that shape our research project:

1. **Corpus creation with the Shared Task on Hope Speech Detection for Equality, Diversity and Inclusion [19] and *UMUCorpusClassifier* [20]**

   The lack of sufficient datasets in languages other than English forces researchers to either modify those existing corpora to accomplish the objective of their research work or create their own resources so that they can analyse language concentrating on particular linguistic phenomena. For our research, we first examined the corpus compiled for the Shared Task on Hope Speech Detection for Equality, Diversity, and Inclusion [19], but as it was focused on the task of hope speech detection, we didn't find enough results to create the first version of our corpus. Consequently, we completed the selection of tweets we found with intention indicators in this corpus with a compilation of tweets extracted through the Twitter API thanks to the extracting tool *UMUCorpusClassifier* [20]. By combining both resources, we were able to compile a corpus of Spanish tweets about the LGTBIQ+ community. The final amount of tweets is 454, which gives us a corpus of 996 instances to analyse, as we decided to tag the intention of each of the

utterances comprised in the same tweet, not the tweet as a whole. We made this decision after noticing that different utterances of the same tweet can show linguistic patterns linked to different intentions, so we preferred to separate the tweets in their utterances to not confuse the recognition of a given intention.

2. **Communicative intentions annotation scheme**

   Parallel to the corpus creation, we compiled linguistic patterns linked to a particular intention according to the SAT classification we explained previously in Section 2. To this end, several resources were also of help to gather the best representation -within our means- of the Spanish linguistic structures that reflect an intention when used appropriately. On the one hand, we translated the verb lexicon comprised in [21] to get a Spanish equivalent of the verbs that, according to Austin and Searle, reflect a particular intention. This book provides in-detailed semantic descriptions of around 200 of the most frequent speech act verbs used in English. In this way, by studying the semantic particularities of each English verb, we could look for the equivalent verbs in Spanish that kept each semantic nuance so that the speech act verb classification would not differ from one language to another. On the other hand, we revisited two of the most extendedly used grammar references of the Spanish language to study their approach to detect speech acts through grammatical features in Spanish [22, 23, 24, 25].

3. **Corpus annotation with INCEpTION [26]**

   The linguistic tool used to annotate our corpus of tweets was the platform created for semantic annotation with intelligent assistance INCEpTION [26]. Thanks to its intuitive interface, this platform allows users to individually manage, curate and modify annotation projects in the same environment by assigning each project to the corresponding annotators. In our case, two experts in the field of Spanish linguistics served as the annotators for our corpus together with the author. As previously mentioned, the tweets were uploaded in the platform and annotated utterance by utterance, as shown in Figure 1, so we could identify as many intention indicators as possible even within the same tweet. Once the annotation task was completed, several metrics were calculated in order to check the inter-annotator agreement achieved between the three annotators. INCEpTION also includes a section that calculates both Fleiss' Kappa and Krippendorf's Kappa automatically in the annotation project, so we checked for the results and confirmed that our annotation scheme could be validated, thanks to achieving a 0.77 of agreement with both measures.

4. **Proposed experiment A): enriching our corpus while training the classification system through active learning**

   Once we validated our annotation scheme, another task we currently study is exploiting the "active learning" functionality that the annotation tool INCEpTION includes within its platform. This machine learning method consists in training the classification system

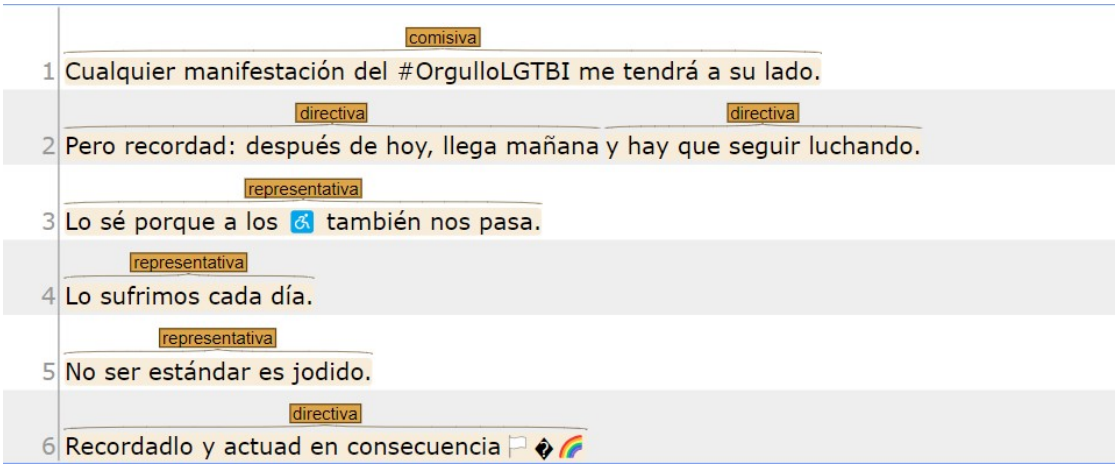

Figure 1: Example of annotating a tweet with intentions in INCEpTION

with the instances we have manually annotated. Then, once we add new instances to be automatically annotated by it, we check which examples the recommender can annotate correctly or not. Those more difficult examples would be the ones to annotate manually so that the recommender system keeps learning on those more ambiguous examples until it finally classifies well those difficult tags without manual help. Consequently, with this technique we will both boost the classifier performance and augment the final weight of our corpus.

5. **Proposed experiment B): combining SAT with sentiments and emotions**

   The second experiment to be fulfilled during our research is combining the identification of communicative intentions in tweets with the tasks of sentiment analysis and emotion detection. Following [12], they demonstrated that previously classifying the sentiment and emotion of a given tweet in English could help to better identify the intention of the tweet. Therefore, as we previously mentioned in the research questions of our doctoral thesis, one of the main tasks we want to accomplish is to check to which point this combined classification can also help to improve the identification of Spanish intentions in tweets. In this way, we would enrich NLP systems with further semantic-pragmatic information and establish more linguistic patterns that help detect the natural essence of a given message. It is also worth mentioning that this research work is being done in collaboration with the Laboratoire Interdisciplinaire des Sciences du Numérique from the Université Paris-Saclay, as one of the research outputs derived from our international Ph.D. stay at the Sémantique et Extraction d'Information Research Group.

6. **Proposed experiment C): incorporating our corpus as a training dataset in a NLG system**

Finally, the last experiment we want to test with our enriched corpus is including it in a NLG architecture as its training dataset so that the system can learn from our already validated examples of messages with a given intention. In this way, we would check if such systems would generate automatic messages with a clear intention following the taxonomy we established in our annotation scheme. To accomplish so, we will follow the methodology established in the task of commonsense text generation, where external knowledge is included as an input in the generation system so that it can generate messages with further world knowledge and linguistic context. In our case, our intention-annotated corpus would be the contextual seed that teaches the system how to recognise an intention, and then generate a new message keeping that same intention, therefore improving the performance of such system by adding more semantic-pragmatic information in its architecture.

5. Research issues to discuss

Given the suggestions and comments received in the previous editions of the Doctoral Symposium, we solved some of the research issues we came up with through the development of our study. However, as an inherent part of this project, new research questions arise that need to be discussed to ensure a good research quality that provides new knowledge within our research area in Spanish:

- Should we add indirect speech act examples to our corpus to check if the classification system is capable of correctly detecting the semantic differences between direct and indirect speech acts?
- Do we have enough intention indicators so that the automatic classification system is capable of differentiating between the different types of intentions we included in our scheme?
- Would it be possible to apply our annotation scheme to other textual typologies outside CMC?
- What if we try to test LLMs ability to generate sentences with a given intention, to check whether there are inconsistencies regarding their intention classification, or if they could be of help to find even more intent linguistic patterns in Spanish?

Acknowledgments

This research work is part of the R&D project "CORTEX: Conscious Natural Text Generation" (PID2021-123956OB-I00), funded by MCIN/AEI/10.13039/501100011033/ and by “ERDF A way of making Europe”.

References

[1] J. L. Austin, How to Do Things with Words, Oxford at the Clarendon Press, 1962.

[2] J. R. Searle, Speech Acts: An Essay in the Philosophy of Language, volume 626, Cambridge University Press, 1969.
[3] J. R. Searle, Expression and meaning: Studies in the theory of speech acts, Cambridge University Press, 1985.
[4] W. C. Mann, Toward a Speech Act Theory for Natural Language Processing, Technical Report, University of Southern California Marina del Rey Information Science Inst, 1980.
[5] S. C. Herring, D. Stein, T. Virtanen, Introduction to the pragmatics of computer-mediated communication, in: Pragmatics of Computer-Mediated Communication, De Gruyter Mouton, 2013, pp. 3–32. doi:10.1515/9783110214468.
[6] C. Bonial, L. Donatelli, M. Abrams, S. Lukin, S. Tratz, M. Marge, R. Artstein, D. Traum, C. Voss, Dialogue-amr: abstract meaning representation for dialogue, in: Proceedings of the 12th Language Resources and Evaluation Conference, 2020, pp. 684–695.
[7] A. Georgakopoulou, Computer-mediated communication, in: J. Verschueren, J.-O. Östman, J. Blommaert, C. Bulcaen (Eds.), Pragmatics in Practice, volume 9, John Benjamins Publishing Co, 2011, pp. 93–110.
[8] T. Saha, S. Saha, P. Bhattacharyya, Tweet act classification: A deep learning based classifier for recognizing speech acts in twitter, in: 2019 International Joint Conference on Neural Networks (IJCNN), IEEE, 2019, pp. 1–8. doi:10.1109/IJCNN.2019.8851805.
[9] R. Zhang, D. Gao, W. Li, What are tweeters doing: Recognizing speech acts in twitter, in: Workshops at the Twenty-Fifth AAAI Conference on Artificial Intelligence, 2011, pp. 86–91. URL: https://www.aaai.org/ocs/index.php/WS/AAAIW11/paper/view/3803.
[10] Y. Tian, T. Galery, G. Dulcinati, E. Molimpakis, C. Sun, Facebook sentiment: Reactions and emojis, in: Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media, ACL, 2017, pp. 11–16. doi:10.18653/v1/W17-1102.
[11] T. Mahler, W. Cheung, M. Elsner, D. King, M.-C. de Marneffe, C. Shain, S. Stevens-Guille, M. White, Breaking NLP: Using morphosyntax, semantics, pragmatics and world knowledge to fool sentiment analysis systems, in: Proceedings of the First Workshop on Building Linguistically Generalizable NLP Systems, Association for Computational Linguistics, Copenhagen, Denmark, 2017, pp. 33–39. URL: https://aclanthology.org/W17-5405. doi:10.18653/v1/W17-5405.
[12] T. Saha, A. Upadhyaya, S. Saha, P. Bhattacharyya, Towards sentiment and emotion aided multi-modal speech act classification in Twitter, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Online, 2021, pp. 5727–5737. URL: https://aclanthology.org/2021.naacl-main.456. doi:10.18653/v1/2021.naacl-main.456.
[13] I. Casanueva, T. Temčinas, D. Gerz, M. Henderson, I. Vulić, Efficient intent detection with dual sentence encoders, in: Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI, Association for Computational Linguistics, Online, 2020, pp. 38–45. URL: https://aclanthology.org/2020.nlp4convai-1.5. doi:10.18653/v1/2020.nlp4convai-1.5.
[14] C. D. Martínez-Hinarejos, J. M. Benedí, V. Tamarit, Unsegmented dialogue act annotation and decoding with n-gram transducers, IEEE/ACM Transactions on Audio, Speech, and Language Processing 23 (2014) 198–211. doi:10.1109/TASLP.2014.2377595.

[15] M. Caballero, L. Díaz, M. Taulé, Guía de anotación del corpus FerroviELE, 2014.
[16] B. Gallardo Paúls, M. Fernández Urquiza, Etiquetado pragmático de datos clínicos, e-AESLA (2015) 1–12.
[17] D. Pascual, Speech acts in travel blogs: Users'corpus-driven pragmatic intentions and discursive realisations, ELIA: Estudios de Lingüística Inglesa Aplicada (2021) 85–123.
[18] S. Ridao Rodrigo, Actos de habla en redes sociales: perfiles privados versus perfiles públicos, Literatura y lingüística (2021) 429–446.
[19] B. R. Chakravarthi, V. Muralidaran, R. Priyadharshini, S. C. Navaneethakrishnan, J. P. McCrae, M. Á. García-Cumbreras, S. M. Jiménez-Zafra, R. Valencia-García, Shared task on hope speech detection for equality, diversity, and inclusion - ACL, 2022. URL: https://competitions.codalab.org/competitions/36393#learn_the_details-organizers.
[20] J. A. García-Díaz, Á. Almela, G. Alcaraz-Mármol, R. Valencia-García, UMUCorpusClassifier: Compilation and evaluation of linguistic corpus for natural language processing tasks, Procesamiento del Lenguaje Natural 65 (2020) 139–142.
[21] A. Wierzbicka, English Speech Act Verbs: A Semantic Dictionary, Academic Press, 1987.
[22] R. A. Española, et al., Nueva gramática de la lengua española, volume 2, Espasa Madrid, 2009.
[23] V. Demonte, Gramática descriptiva de la lengua española: Sintaxis básica de las clases de palabras, volume 1, Espasa, 1999.
[24] V. Demonte, Gramática descriptiva de la lengua española: Las construcciones sintácticas fundamentales, volume 2, Espasa, 1999.
[25] I. Bosque, Gramática descriptiva de la lengua española: Entre la oración y el discurso. Morfología, volume 3, Espasa, 1999.
[26] J.-C. Klie, M. Bugert, B. Boullosa, R. E. de Castilho, I. Gurevych, The INCEpTION platform: Machine-assisted and knowledge-oriented interactive annotation, in: Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations, Association for Computational Linguistics, 2018, pp. 5–9. URL: http://tubiblio.ulb.tu-darmstadt.de/106270/.