

Annotating and normalizing biomedical NEs with limited knowledge*

Fernando Sánchez León
unaffiliated

f.sanchez.lcmcvp@gmail.com

Ana González Ledesma
unaffiliated

ana.gonzalez.ledesma@protonmail.com

Abstract

Named entity recognition (NER) is the very first step in the linguistic processing of any new domain. It is currently a common process in BioNLP on English clinical text. However, it is still in its infancy in other major languages, as it is the case for Spanish. Presented under the umbrella of the PHARMAConER shared task, this paper describes a very simple method for the annotation and normalization of pharmacological, chemical and, ultimately, biomedical named entities in clinical cases. The system developed for the shared task is based on limited knowledge, collected, structured and munged in a way that clearly outperforms scores obtained by similar dictionary-based systems for English in the past. Along with this recovering of the knowledge-based methods for NER in sub-domains, the paper also highlights the key contribution of *resource-based* systems in the validation and consolidation of both the annotation guidelines and the human annotation practices. In this sense, some of the authors discoverings on the overall quality of human annotated datasets question the above-mentioned ‘official’ results obtained by this system, that ranked second (0.91 F1-score) and first (0.916 F1-score), respectively, in the two PHARMAConER subtasks.

1 Introduction

Named Entity Recognition (NER) is considered a necessary first step in the linguistic processing of any new domain, as it facilitates the development of applications showing co-occurrences of domain entities, cause-effect relations among them, and,

This paper should have been published in the *Proceedings of the 5th Workshop on BioNLP Shared Tasks*. Unfortunately, due to their complete lack of funding, the authors could not afford the registration fees, a mandatory expense for a contribution to be published in the aforementioned proceedings.

eventually, it opens the (still to be reached) possibility of understanding full text content. On the other hand, Biomedical literature and, more specifically, clinical texts, show a number of features as regards NER that pose a challenge to NLP researchers (Cohen and Demner-Fushman, 2014): (1) the clinical discourse is characterized by being conceptually very dense; (2) the number of different classes for NEs is greater than traditional classes used with, for instance, newswire text; (3) they show a high formal variability for NEs (actually, it is rare to find entities in their “canonical form”); and, (4) this text type contains a great number of ortho-typographic errors, due mainly to time constraints when drafted.

Many ways to approach NER for biomedical literature have been proposed, but they roughly fall into three main categories: rule-based, dictionary-based (sometimes called knowledge-based) and machine-learning based solutions. Traditionally, the first two approaches have been the choice before the availability of Human Annotated Datasets (HAD), albeit rule-based approaches require (usually hand-crafted) rules to identify terms in the text, while dictionary-based approaches tend to miss medical terms not mentioned in the system dictionary (Rehbolz-Schumann et al., 2011). Nonetheless, with the creation and distribution of HAD as well as the development and success of supervised machine learning methods, a plethora of data-driven approaches have emerged —from Hidden Markov Models (HMMs) (Ephraim, 2002), Support Vector Machines (SVMs) (Habib and Kalita, 2010) and Conditional Random Fields (CRFs) (He and Kayaalp, 2008), to, more recently, those founded on neural networks (Armengol-Estepé et al., 2019). This fact has had an impact on knowledge-based methods, demoting them to a second plane. Besides, this situation has been favoured by claims on the

uselessness of gazetteers for NER in, for example, Genomic Medicine (GM), as it was suggested by Cohen and Demner-Fushman (2014, p. 26):

One of the findings of the first BioCreative shared task was the demonstration of the long-suspected fact that gazetteers are typically of little use in GM.

Although one might think that this view could strictly refer to the subdomain of GM and to the past —BioCreative I was a shared task held back in 2004—, we can still find similar claims today, not only referred to rule-based and dictionary-based methods, but also to stochastic ones (Armengol-Estabé et al., 2019).

In this paper, in spite of previous statements, we present a system that uses rule-based and dictionary-based methods combined (in a way we prefer to call *resource-based*). Our final goals in the paper are two-fold: on the one hand, to describe our system, developed for the PHARMACoNER shared task¹, dealing with the annotation of some of the NEs in health records (namely, pharmacological, chemical and biomedical entities) using a revisited version of rule- and dictionary-based approaches; and, on the other hand, to give pause for thought about the quality of datasets (and, thus, the fairness) with which systems of this type are evaluated, and to highlight the key role of resource-based systems in the validation and consolidation of both the annotation guidelines and the human annotation practices.

In section 2, we describe our initial resources and explain how they were built, and try to address the issues posed by features (1) and (2) above. Section 3 depicts the core of our system and the methods we have devised to deal with text features (3) and (4). Results obtained in PHARMACoNER by our system are presented in section 4. Section 5 details some of our errors, but, most importantly, focusses on the errors and inconsistencies found in the evaluation dataset, given that they may shed doubts on the scores obtained by any system in the competition. Finally, we present some concluding remarks in section 6.

¹<http://temu.bsc.es/pharmaconer/>

2 Resource building

As it is common in resource-based system development, special effort has been devoted to the creation of the set of resources used by the system. These are mainly two —a flat subset of the SNOMED CT medical ontology², and the library and a part of the contextual regexp grammars developed by Sánchez-León (2018) for a previous competition on abbreviation resolution in clinical texts written in Spanish. The process of creation and/or adaptation of these resources is described in this section.

2.1 SNOMED CT

Although the competition proposes two different scenarios, in fact, both are guided by the SNOMED CT ontology —for subtask 1, entities must be identified with offsets and mapped to a predefined set of four classes (PROTEINAS, NORMALIZABLES, NO_NORMALIZABLES and UNCLEAR); for subtask 2, a list of all SNOMED CT IDs (sctid) for entities occurring in the text must be given, which has been called *concept indexing* by the shared task organizers³. Moreover, PHARMACoNER organizers decided to promote SNOMED CT substance IDs over product, procedure or other possible interpretations also available in this medical ontology for a given entity. This selection must be done even if the context clearly refers to a different concept, according to the annotation guidelines⁴ (henceforth, *AnnotGuide*) and the praxis. Finally, PROTEINAS is ranked as the first choice for substances in this category.

These previous decisions alone on the part of the organizers greatly simplify the task at hand, making it possible to build (carefully compiled) subsets of the entities to be annotated. This is a great advantage over open domain NER, where (like in GM) the texts may contain an infinite (and very creative indeed) number of NEs. For clinical cases, although the NE density is greater, there exist highly structured terminological resources for the domain. Moreover, the set of classes to use

²From <https://browser.ihtsdotools.org/>.

³In the train+dev datasets, only 17 of the PROTEINAS ('proteins') and NORMALIZABLES ('standardizable') entities have an ID not in the SNOMED CT ontology. Besides, just 40 out of 5,615 annotations —not taking into account the class UNCLEAR, which is not considered for the system evaluation— are tagged as NO_NORMALIZABLES ('non standardizable'), many of them due to the fact that they include elliptical constructions.

⁴<https://bit.ly/2qxofgd>, p. 4.

in the annotation exercise for subtask 1 has been dramatically cut down by the organizers.

With the above-mentioned initial constraints in mind, we have painstakingly collected, from the whole set of SNOMED CT terms, instances of entities as classified by the human annotators in the datasets released by the organizers and, when browsing the SNOMED CT web version, we have tried to use the ontological hierarchical relations to pull a complete class down from SNOMED CT. This way, we have gathered 80 classes — from lipids to proteins to peptides or peptide hormones, from plasminogen activators to dyes to drugs or medicaments—, that have been arranged in a ranked way so as to mimic human annotators choices⁵. The number of entities so collected (henceforth, ‘primary entities’) is 51,309.

2.2 Contextual regexp grammars

Some of the entities to be annotated, specially those in abbreviated form, are ambiguous without a context. This is the case, for instance, of *PCR*, whose expanded forms are (among other meanings; we use only English expanded forms) ‘reactive protein c’, ‘polymerase chain reaction’, ‘cardiorespiratory arrest’. In order to deal with these cases, we use a contextual regexp rule system with a lean and simple rule formalism previously developed (Sánchez-León, 2018). As an exemplification, we include one rule to deal with one of the cases of the preceding ambiguity:

```
b:[il:::bioquímica|en sangre|hemoglobina|  
hemograma|leucocit|parásito|plaquetal|  
prote.na|recuento|urea] - [PCR] ->  
[m=proteína]
```

A rule has a left hand side (LHS) and a right hand side (RHS). There is a focus in the LHS (*PCR*, within dashes) and a left and right context (that may be empty). When the left context includes a *b*: (like in this case), it indicates either left or right context. The words in the context can take other qualifiers —in this case, the matching will be case insensitive (*i* to the left of *bioquímica*) and local (*l*), which means the disjunction of words and/or stems can be found in a distance of 40 characters (this can be modified by the user). Hence, the rule applies, selecting the *proteína* expansion (in RHS) of *PCR* if any of the words/stems specified as local context (40

⁵Note that we have gathered the complete set of medical terms included in SNOMED CT, but, for the purpose of this shared task, we only use a subset of it.

chars maximum) is matched either to the left or right of the focus term (which is usually an abbreviation).

With no tweaking at all for the datasets in PHARMACONER competition, the system annotates correctly 18 out of 20 occurrences of *PCR* in the test dataset (a precision of 0.9)⁶.

This component of the system is important because, only when the previous abbreviation is expanded as the first string (that of a protein name), it must be annotated, according to the *AnnotGuide*. The same ambiguity happens with *Cr*, which may mean ‘creatinine’ or ‘chrome’⁷. These expansions are both NORMALIZABLES, but, obviously, their sctid is different.

The system currently uses 104 context rules, only for abbreviations and acronyms in the clinical cases. These rules, contrary to what is commonly referred in the biomedical processing literature (Armengol-Estabé et al., 2019), do not require a special domain knowledge (none of the authors do have it) and can be written, most of the times, in a very straightforward way in the formalism briefly described above.

3 Development

In general, dictionary-based methods rely on strict string matching over a fixed set of lexical entries from the domain. This is clearly insufficient to deal with non-canonical linguistic forms of NEs as used in clinical texts. For this reason, we have devised two different solutions to this shortcoming.

In the first place, we have munged a great number of our primary entities, in a way similar to that described in Sánchez-León (2019) for gazetteers used for protected information anonymization in clinical texts. We basically transform canonical forms in other possible textual forms observed when working with biomedical texts. With such transformations, a system module converts a salt compound like *clorhidrato de ciclopentolato* into *ciclopentolato clorhidrato*, or simply the *PP de potasio* into its corresponding adjective *potásico*. Other, more complex conversions include the treatment of antibodies —for instance *anticuerpo contra especie de Leishmania* becomes *ac. Leish-*

⁶Note that 2 of the *PCR* occurrences in the train+dev datasets have been incorrectly mapped to the protein interpretation (file S1130-63432014000100012-1, 2 times).

⁷Again, one of the occurrences of *Cr* has been incorrectly mapped to the former extended form (file S0212-16112012000500042-1).

mania, among other variants—, or pairs of antibiotics normally prescribed together —which have a unique scid and whose order we handle just as the ‘glueing’ characters. Note, incidentally, that, while the input to this pre-processing step is always a string, the output can be a regular expression, that is linked to a scid. Plural forms are also generated through this module, that uses 45 transformations (not all equally productive). Using these transformation rules, we produce 139,150 ‘secondary entities’, many of them regexps. As a final (simple) example of this, consider the entity *antígeno CD13*: after applying one of the previous string-to-regexp transformations, it is converted to:

```
(?:antígeno )?CD[- ]?13
```

With the previous regexp, the system is able to identify (and string-normalize) six different textual realizations of the same unique SNOMED CT term. There are more complex rules that, thus, produce many more potential strings. The important thing with this strategy is that through the generative power of these predictably-created regexps from SNOMED CT entities the system is able to improve its recall and overcome the limitations of traditional dictionary-based approaches.

Secondly, to tackle with careless drafting of clinical reports, a Levenshtein edit distance library⁸ is used on the whole background dataset. The process is run once, using our secondary entities as lexicon⁹ and a general vocabulary lexicon to rule out common words in the candidate search process. We have used distances in the range 1-3 (depending on string length) for sequences up to

⁸We use `Text::Levenshtein::Flexible` library, from Perl ecosystem. One of the anonymous reviewers has shed doubts about the use of Perl as a language for “NLP and text-mining nowadays”. In this respect, we are not committed with a given programming language more than we are with our native language —and we have submitted our paper in English, a foreign language for us. The system could have been implemented in any other programming language more popular “nowadays”, provided that we were as *proficient* in it as we are in Perl and the language used were as efficient in string and regexp handling and in I/O operations as Perl is. In this regard, the most popular language nowadays —Python— is 2 to 10 times slower for these particular features. Perl is even faster for regexp processing than Python PyPy —see, for instance, <https://github.com/mariomka/regex-benchmark>. Idiomatic Perl is even faster. Finally, Perl has a long tradition in biology and medicine text processing.

⁹With enumeration of strings from non-infinite-loop regexps.

3 words long¹⁰. The output of this process, which links forms with spelling errors with canonical ones and, thus, to scids, can be inspected prior to its inclusion in the system lexicon, if so desired.

3.1 Annotation process

As such, the annotation process is very simple. The program reads the input byte stream trying to identify known entities by means of a huge regexp built through the pre-processing of the available resources. If the candidate entity is ambiguous and (at least) one contextual rule exists for it, it is applied. For the rest of the NEs, the system assigns them the class and scid found in our ranked in-memory lexicon. As already mentioned in passing, the system does not tokenize text prior to NER, a processing order that we consider the right choice for highly entity-dense texts. The data structures built during pre-processing are efficiently stored on disk for subsequent runs, so the pre-processing is redone only when resources are edited.

4 Results

According to the organizers, and taking into account the HA of the tiny subset from the background dataset released to the participants¹¹, the system obtained the scores presented in table 1, ranking as second best system for subtask1 and best system for subtask2 (Gonzalez-Agirre et al., 2019).¹²

Our results are consistent with our poor understanding of the classes for subtask 1. Having a null knowledge of Pharmacology, Biomedicine or even Chemistry, assigning classes (as requested for subtask 1) to entities is very hard, while

¹⁰These words are not isolated from the byte stream, and the process uses textual anchors to delimit them as word candidates. Consequently, no proper tokenization is performed.

¹¹When compared with the rest of the tasks in BioNLP-OST 2019, the time given to PHARMACoNER participants to submit their system runs is 4 times longer than the mean —longer time that is unnecessary if system is mature enough. On the other hand, the dataset released for evaluation purposes is more than 4 times larger than the mean. As a consequence, participating groups have to annotate full domain corpora rather than just test dataset(s). A shorter submission period and a smaller test dataset would be preferable, and besides fairer, in future calls.

¹²The authors have been unable to obtain these results with the official script, downloaded from <https://github.com/PlanTL-SANIDAD/PharmaCoNER-CODALAB>. In their execution of the evaluation script, system results are better (?).

| | Precision | Recall | F1-score |
|------------------|-----------|---------|----------|
| Subtask 1 | 0.90625 | 0.91314 | 0.90968 |
| Subtask 2 | 0.91108 | 0.92083 | 0.91593 |

Table 1: Results for PHARMAConER test dataset (both subtasks)

providing a scid (subtask 2) seems an easier goal. We will explain the point with an example entity —*ácido hialurónico* (‘hyaluronic acid’). Using the ontological structure of SNOMED CT, one can find the following parent relations (just in English):

hyaluronic acid IS-A *mucopolysaccharide*
IS-A *protein*

The authors have, in this case, promoted the PROTEINAS annotation for this entity, disregarding its interpretation as a replacement agent and overlooking a recommendation on polysaccharides in the *AnnotGuide*. Fortunately, all its interpretations share a unique scid. The same may be true for

haemosiderin IS-A *protein*

which is considered NORMALIZABLE in the test dataset. Similar cases are responsible for the lower performance on subtask 1 with respect to the more complex subtask 2.

In spite of these human classification errors, our system scores outperform those obtained by PharmacoNER Tagger¹³ (Armengol-Estabé et al., 2019), a simpler system using a binary classification and a very different organization of the dataset with a smaller fragment for test (10% of the data as opposed to 25% for the official competition). In fact, our system improves their F1-score (89.06) by 1.3 points when compared with our results for the more complex PHARMAConER subtask 1.

5 Discussion

In this section, we perform error analysis for our system run on the test dataset. We will address both recall and precision errors, but mainly concentrate on the latter type, and on a thorough

¹³The tagger authors, some of them also organizers of shared task, have changed the casing of the name for the program.

revision of mismatches between system and human annotations.

In general, error analysis is favoured by knowledge-based methods, since it is through the understanding of the underlying reasons for an error that the system could be improved. Moreover, and differently to what happens with the current wave of artificial neural network methods, the whole annotation process —its guidelines for human annotators, the collection and appropriate structuring of resources, the adequate means to assign tags to certain entities but not to other, similar or even pertaining to the same class— must be clearly understood by the designer/developer/data architect of such systems. As a natural consequence of this attempt to mimick a task defined by humans to be performed, in the first place, also by humans, some inconsistencies, asystematic or missing assignments can be discovered, and this information is a valuable treasure not only for system developers but also for task organizers, guideline editors and future annotation campaigns, not to mention for the exactness of program evaluation results.

Most of the error types made by the system (i.e., by the authors) in class assignment for subtask 1 have already been discussed. In the same vein, as regards subtask 2, a great number of errors come from the selection of the ‘product containing substance’ reading from SNOMED CT rather to the ‘substance’ itself. This is due to inexperience of the authors on the domain and the wrong consideration of context when tagging entities —the latter being clearly obviated in the *AnnotGuide*.

In the following paragraphs, some of the most relevant inconsistencies found when performing error analysis of our system are highlighted. The list is necessarily incomplete due to space constraints, and it is geared towards the explanation of our possible errors.

5.1 Inconsistency in the AG

Among some of the paradoxical examples in the *AnnotGuide* it stands out the double explicit consideration of *gen* (‘gene’), when occurs alone in context, as both an entity to be tagged (positive rule P2 of the *AnnotGuide*) and a noun not to be tagged (negative rule N2). This inconsistency (and a bit of bad luck) has produced that none of the 6 occurrences as an independent noun —not in-

troducing an entity—is tagged in the train+dev (henceforth, t+d) while the only 2 in the same context in the test dataset have been tagged. This amounts for 2 true negatives (TNS) for the evaluation script.

5.2 Inconsistency in HA as regards AG

The *AnnotGuide* proposal for the treatment of elliptical elements is somewhat confusing. For these cases, a longest match annotation is proposed, which is difficult to replicate automatically and not easy to remember for the human annotator. In many contexts, the annotator has made the right choice—for instance, in *receptores de estrógeno y de progesterona*—whereas in others do not—*|anticuerpos anticardiolipina| IgG e IgM*, with ‘|’ marking the edges of the annotations. The last example occurs twice in the test dataset. Hence, the disagreement counts as 6 TNS and 2 false positives (FPS)¹⁴.

On the other hand, there is a clear reference to food materials and nutrition in the *AnnotGuide*, where they are included in the class of substances. However, none of the following entities is tagged in the test dataset: *azúcar* (which is mandatory according to *AnnotGuide* and was tagged in t+d; 1 FP); *almidón de maíz* (also mandatory in *AnnotGuide*; 1 FP); and *Loprofín, Aglutella, Aprotén* (hypoproteic nutrition products, 3 FPS in total)¹⁵.

There is an explicit indication in the *AnnotGuide* to annotate salts, with the example *iron salts*. However, in the context *sales de litio* ('lithium salts'), only the chemical element has been tagged (1 FP)¹⁶.

There exist other differing-span mismatches between human and automatic annotation. These include *anticuerpos anticitoplasma de neutrófilo*, where the HA considers the first two words only (in one of the occurrences, 1 FP); in the text fragment *b2 microglobulina, CEA y CA 19,9 normales, CA 19,9* is the correct span for the last entity (and not *CA*, 1 FP); *A.S.T* is the span selected (for *A.S.T.*, 1 FP); finally, in the context *lgM anticore* only *lgM* has been tagged (1 FP).

Other prominent mismatch between HAD and *AnnotGuide* is that of *DNA*, which is explicitly

¹⁴When we indicate this kind of information, mostly using only FPS, it must be understood that the system made the choice(s) that the authors judge as correct, although disagreeing with HA and/or *AnnotGuide*.

¹⁵On nutrition replacements, see also section 5.3.

¹⁶Note, in passing, that these span errors account for 1 TN also for the evaluation scripts.

included in the *AnnotGuide* (sects. P2 and O1). It accounts for 2 FPS.

But perhaps one of the most common discrepancies between human and automatic annotation has to do with medicaments normally prescribed together, which have a unique scid. Examples include *amiloride/hidroclorotiazida* (1 FP); and *betametasona + calcipotriol* (1 FP) in the test set. This situation was also observed in the t+d corpus fragment (*tenofovir + emtricitabina, carbonato cálcico /colecalciferol, lopinavir/ritonavir*).

5.3 Inconsistency in HA on the test set as regards t+d sets

Some inconsistencies between dataset annotations have turned the authors crazy: *NPT* (acronym for ‘total parenteral nutrition, TPN’) is tagged in the train+dev dataset 15 out of 21 times it occurs¹⁷. The common sense of frequency in the HA of texts has led us to tag it in the background set. Unluckily, neither *NPT* nor its expansion have been tagged in the test dataset. This has also been the behaviour in HA for ‘parenteral nutrition’ and ‘enteral nutrition’ (and their corresponding acronyms) in test dataset, since these entities have not been tagged. We asked the organizers about this and other entities for which we had doubts, either because the *AnnotGuide* didn’t cover their cases or because the HA didn’t match the recommendations in the *AnnotGuide*. Woefully, communication with the organizers has not been very fluent on this respect. All in all, this bad decision on the part of the authors amounts for 6 FPS (more than 7.5% of our FPS according to evaluation script).

For other cases, decisions that may be clearly induced from the tagging of train+dev datasets, have not been applied in the test corpus fragment. These include *cadenas ligeras* (5 times in t+d, 1 FP in test); *enzimas hepáticas* (tagged systematically in t+d, 1 FP); *p53* (also tagged in t+d, 1 FP).

Another entity that stands out is *hidratos de carbono* ('carbohydrates'). It is tagged twice in the t+d dataset, occurring 4 times in the set (once as *HC*). However, although the form *carbohidratos* has been annotated twice in the test set, *hidratos de carbono* has been not (1 FP).

Moreover, *suero* ('Sodium chloride solution' or

¹⁷However, at least one expanded variant of it—*nutrición parenteral*, ‘parenteral nutrition’—is never tagged.

‘serum’) deserves its own comment. Both entity references are tagged in the train+dev datasets (although with the latter meaning it is tagged only 4 out of 12 occurrences). We decided to tag it due to its relevance. In the test dataset, it occurs 5 times with the blood material meaning, but it has only been tagged twice as such (one of them being an error, since it refers to the former meaning). Our system tagged all occurrences, but tagged also one of the instances with the former meaning as serum (3 FPs).

Finally, there are some inconsistencies within the same dataset. For example, nutricional agent *Kabiven* is tagged as both NORMALIZABLES (with scid) and NO_NORMALIZABLES in the very same text. The same happens with another nutritional complement, *Cernebit*, this time in two different files. The perfusion solution *Isoplasmal G* (with a typo in the datasets —*Isoplasmar G*) is tagged as NORMALIZABLES and UNCLEAR. These examples reveal a vague understanding (or definition) of criteria as regards fluids and nutrition, as we pointed out at the beginning of this section.

5.4 Asystematic/incomplete annotation

Some of the entities occurring in the test dataset have not always been tagged. This is the case for *celulosa* (annotated only once but used twice, 1 FP); *vimentina* (same situation as previous, 1 FP); *LDH* (tagged 20 times in t+d but not in one of the files, 1 FP); *cimetidina* (1 FP); *reactantes de fase aguda* (2 FPs; 2 other occurrences were tagged); *anticuerpos antinucleares* (human annotators missed 1, considered FP).

5.5 Incorrect scids

On our refinement work with the system, some incorrect scids have emerged. These errors impact on subtask 2 (some also on subtask 1). A large sample of them is enumerated below.

ARP (‘actividad de renina plasmática’, ‘plasma renin activity’, PRA) cannot be linked to scid for *renina*, which happens twice. In the context ‘perfil de antígenos [sic] extraíbles del núcleo (ENA)’, ENA has been tagged with scid of the antibody (1 FP). In one of the files, *tioflavina* is linked to scid of *tioflavina T*, but it could be *tioflavina S*. Thus, it should be NO_NORMALIZABLE. *Harvoni* is ChEBI:85082 and not <null> (1 FP). *AcIgM contra CMV* has a wrong scid (1 FP). *HBsAg* has no scid in the test set; it should be 22290004

(‘Hepatitis B surface antigen’) (1 FP).

There are other incorrect annotations, due to inadvertent human errors, like *biotina* tagged as PROTEINAS or *VEB* (‘Epstein-Barr virus’) being annotated when it is not a substance. Among these mismatches between HA and system annotation, the most remarkable is the case of synonyms in active principles. For instance, the brand name drug *Dekapine* has been linked to ‘ácido valproico’ in the former case and to ‘valproato sódico’ in the latter. These terms are synonymous¹⁸, but sadly they don’t share scid. Hence, this case also counts as a FP.

A gold standard dataset for any task is very hard to develop, so a continuous editing of it is a must¹⁹. In this discussion, we have focused on false positives (FPs) according to the script used for system evaluation, with the main purpose of *understanding* the domain knowledge encoded in the linguistic conventions (lexical/terminological items and constructions) used by health professionals, but also the decisions underlying both the *AnnotGuide* and the HA practice.

In this journey to system improvement and authors enlightenment, some inconsistencies, errors, omissions have come up, as it has been reflected in this section, so both the guidelines for and the practice of annotation can also be improved in future use scenarios of the clinical case corpus built and maintained by the shared task organizers.

Our conclusion on this state of affairs is that some of the inconsistencies spotted in this section show that there were not a rational approach to the annotation of certain entities contained in the datasets (apart from other errors and/or oversights), and, hence, the upper bound of any tagging system is far below the ideal 1.0 F1-score. To this respect, in very many cases, the authors have made the wrong choice, but in others they were guided by analogy or common sense. Maybe a selection founded on probability measures estimated on training material could have obtained better results with this specific test dataset. However, in the

¹⁸Although, ‘valproato sódico’ is the name used in the leaflet, as it can be seen in the Spanish Medicament Agency, AEMPS, web page (https://cima.aemps.es/cima/dochtml/p/48828/P_48828.html) last consulted on 16.07.2019.

¹⁹Besides, when the dataset is being used in a shared task, this refinement process should be available to participants while the task is open.

end, this cannot be considered as an indication of a better system performance, since, as it has been shown, the test dataset used still needs more refinement work to be used as the right dataset for automatic annotation evaluation.

6 Conclusions

With this resource-based system developed for the PHARMAConER shared task on NER of pharmacological, chemical and biomedical entities, we have demonstrated that, having a very limited knowledge of the domain, and, thus, making wrong choices many times in the creation of resources for the tasks at hand, but being more flexible with the matching mechanisms, a simple-design system can outperform a NER tagger for biomedical entities based on state-of-the-art artificial neural network technology. Thus, knowledge-based methods stand on their own merits in task resolution.

But, perhaps most importantly, the other key point brought to light in this contribution is that a resource-based approach also favours a more critical stance on the dataset(s) used to evaluate system performance. With these methods, system development can go hand in hand with dataset refinement in a virtuous circle that let us think that maybe next time we are planning to add a new gazetteer or word embedding to our system in order to try to improve system performance, we should first look at our data and, like King Midas, turn our Human Annotated Dataset into a true Gold Standard Dataset.

Acknowledgements

We thank three anonymous reviewers of our manuscript for their careful reading and their many insightful comments and suggestions. We have made our best in providing a revised version of the manuscript that reflects their suggestions. Any remaining errors are our own responsibility.

References

- J. Armengol-Estabé, F. Soares, M. Marimon, and M. Krallinger. 2019. PharmacCoNER Tagger: a deep learning-based tool for automatically finding chemicals and drugs in Spanish medical texts. *Genomics & Informatics*, 17(1):e15.
- K. B. Cohen and D. Demner-Fushman. 2014. *Biomedical Natural Language Processing*. John Benjamins Publishing Company.

- Y.M.N. Ephraim. 2002. Hidden Markov processes. *IEEE Trans Inform Theory*, (48):1518–69.
- A. Gonzalez-Agirre, M. Marimon, A. Intxaurrendo, O. Rabal, M. Villegas, and M. Kralliger. 2019. PharmaCoNER: Pharmacological Substances, Compounds and proteins Named Entity Recognition track. In *Proceedings of The 5th Workshop on BioNLP Open Shared Tasks*, pages 1–10.
- M.S. Habib and J. Kalita. 2010. Scalable biomedical Named Entity Recognition investigation of a database-supported SVM approach. *Int J Bioinform Res Appl*, (6):191–208.
- Y. He and J. Kayaalp. 2008. Biological entity recognition with conditional random fields. In *AMMIA annu symp proc*, pages 293–7.
- D. Rebholz-Schumann, A.J. Yepes, C. Li, S. Kafkas, I. Lewin, and N. Kan. 2011. Assessment of NER solutions against the first and second CALBC Silver Standard Corpus. *J Biomed Semantics*, 2(Supl. 5)(S11).
- F. Sánchez-León. 2018. ARBOREx: Abbreviation Resolution Based on Regular Expressions for BARR2. In *IberEval@SEPLN*, pages 302–315.
- F. Sánchez-León. 2019. Resource-based anonymization for Spanish clinical cases. In *IberLef@SEPLN*, pages 704–711.