Use of unsupervised word classes for entity recognition: Application to the detection of disorders in clinical reports

Maria Evangelia Chatzimina Cyril Grouin Pierre Zweigenbaum
CNRS, UPR 3251, LIMSI
91403 Orsay, France
hatziminamaria@gmail.com, {firstname.lastname}@limsi.fr

Abstract

Unsupervised word classes induced from unannotated text corpora are increasingly used to help tasks addressed by supervised classification, such as standard named entity detection. This paper studies the contribution of unsupervised word classes to a medical entity detection task with two specific objectives: How do unsupervised word classes compare to available knowledge-based semantic classes? Does syntactic information help produce unsupervised word classes with better properties? We design and test two syntax-based methods to produce word classes: one applies the Brown clustering algorithm to syntactic dependencies, the other collects latent categories created by a PCFG-LA parser. When added to non-semantic features, knowledge-based semantic classes gain 7.28 points of F-measure. In the same context, basic unsupervised word classes gain 4.16pt, reaching 60% of the contribution of knowledge-based semantic classes and outperforming Wikipedia, and adding PCFG-LA unsupervised word classes gain one more point at 5.11pt, reaching 70%. Unsupervised word classes could therefore provide a useful semantic back-off in domains where no knowledge-based semantic classes are available. The combination of both knowledge-based and basic unsupervised classes gains 8.33pt. Therefore, unsupervised classes are still useful even when rich knowledge-based classes exist.

Keywords: Clinical Texts; Natural Language Processing; Unsupervised Word Classes

1. Introduction

Clinical texts and biomedical literature are important sources of medical knowledge. The exponentially growing amount of biomedical sources in recent years requires dedicated tools to process this domain (Meystre et al., 2008).

Named entity recognition is the task that identifies an entity's boundaries within text and assigns the entity to their corresponding class or category. It is usually one of the basic steps applied when extracting information from texts. Supervised learning based on annotated corpora, typically using sequence learning algorithms such as Conditional Random Fields (Lafferty et al., 2001; Sutton and McCallum, 2006), is now a standard method to detect entities.

Unannotated corpora are larger than annotated corpora, making them useful for unsupervised methods to learn word classes and help supervised methods perform entity detection (Turian et al., 2010).

The main objective of this work is to study the use of unsupervised word classes in medical entity recognition, and more specifically to test different methods and variants of unsupervised word class construction which use more initial information thanks to syntactic analysis.

The remainder of the paper is organized as follows. Section 2. provides a brief description of previous work related to our study. Section 3. describes the corpora we take as a testbed for our experiments. Section 4. describes the design and implementation of our study. Section 5. presents and discusses experimental results. We conclude in Section 6.

2. Related Work

In an early attempt to help supervised learning with the use of supervised word classes, Brown et al. (1992)'s clustering algorithm was used to detect standard named entities (Miller et al., 2004). The combined method achieved a 25% reduction in error on a standard named-entity problem. In further experiments, Turian et al. (2010) also tested the word embeddings obtained by neural language models (Collobert and Weston, 2008; Mnih and Hinton, 2009) but observed that Brown clusters were superior when helping chunking or named entity recognition.

Using only one type of word representation, Brown clusters had the highest F-measure on the test set compared to Collobert and Weston (2008)'s embeddings and the hierarchical log-bilinear model of Mnih and Hinton (2009), and even outperformed the use of gazetteers by 1 point of F-measure. Combinations of these unsupervised word representations gained close to another point of F-measure.

In the biomedical domain, Brown clusters have been used on top of domain-specific, knowledge-based semantic classes such as the UMLS semantic types (Bodenreider, 2004) for medical entity recognition and relation detection (de Bruijn et al., 2011). The specific contribution of Brown clusters to this top-performing system in the i2b2 2010 challenge (Uzuner et al., 2011) amounted to 0.14 points of F-measure for entity recognition. Jonnalagadda et al. (2012) computed a thesaurus of the 20 distributionally most similar words for each input word and obtained a 2pt increase in F-measure over the same data set. Tang et al. (2013) tested Brown clusters and also the distributionally most similar words as features to help medical entity recognition on the i2b2 2010 data set. They obtained increases of 0.40-0.56pt F-measure in a system which performs at the same level as that of de Bruijn et al. (2011), i.e., starting from a stronger baseline than Jonnalagadda et al. (2012). Jonnalagadda et al. (2013) used distributional representations of words built with the Semantic Vectors package of Widdows and Cohen (2010) to create word clusters, a 'quasi-lexicon', and a thesaurus as above (20 nearest neighbors): the thesaurus improves entity recognition on the i2b2 2010 data set by 1.6pt F-measure, whereas the quasi-lexicon and clusters obtain smaller gains. They also observed that a combination of distributional word classes

(quasi-lexicon and thesaurus) could replace UMLS-based word classes with no loss (and even an increase) of performance.
The present work also aims to assess the relative contribution of unsupervised word classes to medical entity detection, compared to knowledge-based semantic classes such as provided by the UMLS. Besides, it tests new word classes which include syntactic information in their construction, which to our knowledge has not been done previously.

3. Corpora

We used two types of corpora:

- The annotated corpora from the ShARe/CLEF eHealth Evaluation Lab (Suominen et al., 2013). The datasets consist of de-identified clinical free-text notes from the MIMIC II database[1] (Saeed et al., 2011) of Intensive Care Unit (ICU) data. The clinical reports were given with stand-off annotations of disorder mention spans and UMLS concept unique identifiers; we do not consider the latter in the present mention detection task, which corresponds to Task 1A of CLEF eHealth.
- A larger unannotated corpus obtained from the MIMIC-II database besides the CLEF eHealth challenge. It contains about 18,000 discharge summaries.

Table 1 provides statistics on the three corpora.

Type	Set	# reports	# words
Annotated	Training	200	94 k
Annotated	Test	100	88 k
Unannotated		18,000	27 M

Table 1: Statistics on corpora

4. Word classes for supervised medical entity detection

4.1. Baseline system features

For supervised medical entity detection, we reuse a pipeline of components (Bodnari et al., 2013) prepared for the CLEF eHealth challenge. The features produced by this pipeline included knowledge-based semantic classes, but no data-driven classes. Supervised classification was performed with the Wapiti CRF toolkit (Lavergne et al., 2010).[2] The baseline system (later called *nosem*) uses the following features:

- Lexical features: token, lemma.
- Morphological features: (i) token containing only upper case letters, (ii) token is a digit, (iii) is capitalized and (iv) is a punctuation.
- Syntactic features: part of speech information extracted with the cTAKES system (Savova et al., 2010).
- Document structure features: document type and section type, extracted with a rule-based method which identifies occurrences of section names in the text (Bodnari et al., 2013).

For most types, a feature is computed based on the previous position in the sentence, current position and next position (for unigram features), and based on the previous+current positions or current+next positions (for bigram features). This is also the case for the semantic classes defined below. Figure 1 provides an overview of the pipeline used to produce semantic features, which we detail below.

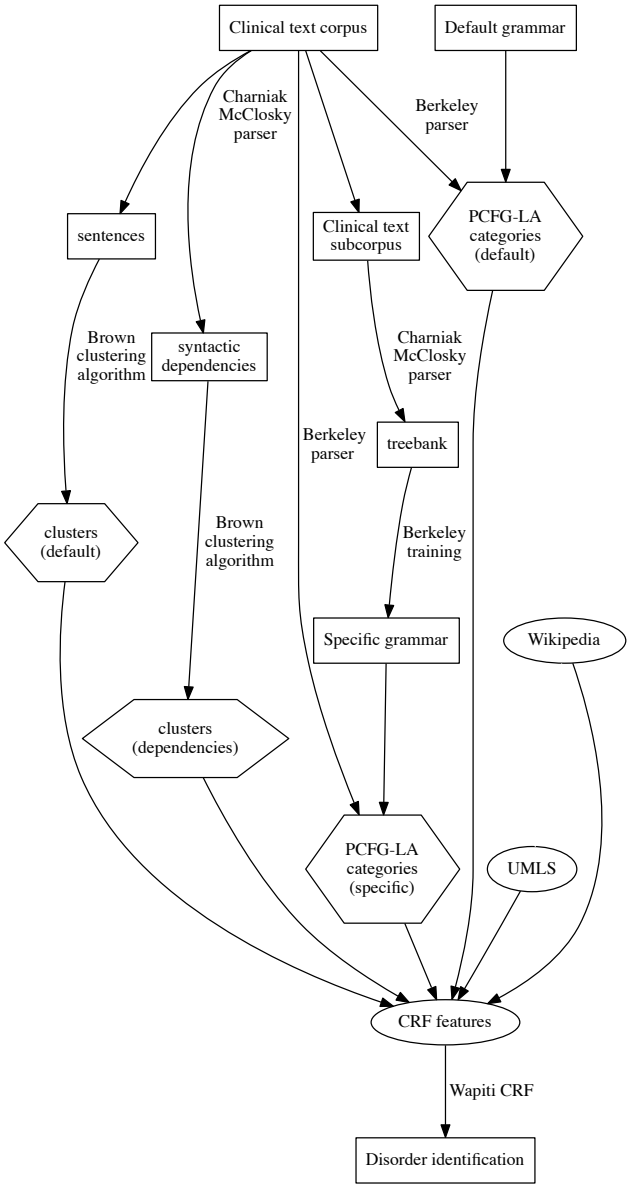

Figure 1: Global overview of semantic feature production

Table 2 shows an excerpt from the corpus in a simplified tabular form, processed by our pipeline. For each token (first column), a set of attributes has been computed. We only show a small subset of these attributes: lemma, part-of-speech, syntactic chunk, the concept unique identifier (CUI) from the UMLS, section from the document (herein, "History of Present Illness"), the cluster ID produced using Brown's algorithm (for a given configuration). We also

[1] http://mimic.physionet.org/
[2] http://wapiti.limsi.fr/

Form	Lemma	POS	Chunk	CUI	Section	Cluster ID	Gold standard	Prediction
The	the	DT	B-NP	#NA	HPI	B-110100	O	O
patient	patient	NN	I-NP	#NA	HPI	B-110001	O	O
is	be	VBZ	B-VP	#NA	HPI	B-01110010	O	O
a	a	DT	O	#NA	HPI	B-010000	O	O
40	40	CD	B-NP	#NA	HPI	B-111101011100	O	O
-	-	HYPH	I-NP	#NA	HPI	B-10010	O	O
year	year	NN	I-NP	#NA	HPI	B-1001111101	O	O
-	-	HYPH	I-NP	#NA	HPI	B-10010	O	O
old	old	JJ	I-NP	#NA	HPI	B-0101110110	O	O
female	female	NN	I-NP	B-C0015780	HPI	B-011000000	O	O
with	with	IN	B-PP	#NA	HPI	B-001111000	O	O
complaints	complaint	NNS	B-NP	B-C0277786	HPI	B-011011111100111	O	O
of	of	IN	B-PP	#NA	HPI	B-001110	O	O
headache	headache	NN	B-NP	B-C0018681	HPI	B-01101001011	B-Disorder	B-Disorder
and	and	CC	I-NP	#NA	HPI	B-00010	O	O
dizziness	dizziness	NN	I-NP	B-C0012833	HPI	B-0110100100	B-Disorder	B-Disorder
.	.	.	O	#NA	HPI	B-1110	O	O

Table 2: Excerpt from the annotated corpus processed by our pipeline. POS = Part-of-Speech; CUI = Concept Unique Identifier; Reference = expected category; Prediction = predicted category

show the gold standard word label and the label predicted by our system.

4.2. Knowledge-based semantic classes

Semantic classes for words were obtained from two human-curated resources (Bodnari et al., 2013).

The UMLS. The Unified Medical Language System (Bodenreider, 2004) contains a large repository of biomedical terms with associated semantic types. We produced for each word eleven attributes based on various ways to access this resource.

The Wikipedia. The Wikipedia contains an increasing number of articles relevant to the medical domain. We used Wikipedia category data to collect a lexicon with nine semantic groups, which was used to produce two attributes for each word.

4.3. Unsupervised, data-driven word classes

Unsupervised learning induces word representations from unannotated corpora. We explore here the use of Brown clusters (Brown et al., 1992) in their usual form as well as a novel application of this method to dependency relations (Section 4.3.1.), then categories obtained through training a probabilistic context-free parser with latent categories (PCFG-LA, (Petrov et al., 2006), Section 4.3.3.).

In all cases, unsupervised learning results in a lexicon which associates words to classes; this lexicon is then applied to the training or test corpus to add attributes to the input words for supervised entity detection.

4.3.1. Brown clusters

Brown clusters optimize a class bigram language model over the corpus on which they are computed, here the unannotated corpus of discharge summaries. We created Brown clusters using Liang (2005)'s implementation of Brown et al. (1992)'s clustering algorithm, requesting 320 and 1000 clusters (these provided the best results in preliminary experiments) combined to five thresholds ($2^{1\ldots5}$) on the minimum number of word occurrences. The 10 resulting clusterings were used as attributes to create CRF features for each input word: for each clustering, the identifier of the cluster which contains this word (or #NA if the word is not present in the clusters) is added as an attribute.

An example of cluster is that named 01101111011111, computed with the unannotated corpus, which contains the following words: *LLL, RLL, RUL, bibasilar, dense, diffuse, expiratory, faint, linear, multifocal, nodular, partial, patchy, retrocardiac, scattered, trace*, etc., many of which occur as modifiers in disorder expressions in the training corpus: *patchy opacities, retrocardiac opacification, expiratory wheeze, inspiratory wheeze*, etc. Among these, *RUL* (right upper lobe) does not occur as a term in the UMLS. In the sentence "*77 year old man with questionable RUL pneumonia*," the word *pneumonia* is present as a disorder in the UMLS, but *RUL* is unknown and as a result the CRF only labels *pneumonia* as a disorder, missing *RUL*. When using Brown clusters, the feature for cluster 01101111011111 obtains a high score for being part of a disorder, and *RUL* is therefore labeled as part of the disorder too, thus obtaining the correct complete entity.

4.3.2. Dependency-based Brown clusters

In previous approaches, the Brown clusters were computed over bigrams in an input string of words. In the present work, we tested whether computing clusters based on syntactic dependencies instead of word bigrams might improve entity detection.

The unannotated corpus was parsed with McClosky and Charniak (2008)'s self-trained parser and the output was converted into Stanford reduced dependency format (De Marneffe and Manning, 2008).

Rather than trying to change the Brown algorithm, we tried to find a representation of the syntactic dependencies that would be amenable to processing directly by the existing

algorithm and implementation. In this purpose, each dependency was represented as a pseudo-sentence, including the two related tokens (governor then dependent) and possibly the grammatical relation. This means that each initial sentence is replaced with a series of pseudo-sentences, one per dependency relation. We tested the following representations of dependency relations: (i) token-relation-token, (ii) relation-token-token, (iii) token-token-relation, and (iv) relation conjoined with the first token (relation-token), second token conjoined with the relation (token-relation) and two tokens without the relation (token-token). Experiments evidenced that the pair of tokens (token-token), without the relation, obtained the best results, so for the sake of space we only consider this setting in this paper.
As for the clusters in Section 4.3.1., for each input word, dependency-based Brown clusters contribute 10 attributes generated by the two numbers of clusters and the five frequency thresholds.

4.3.3. Latent syntactic categories

We also tested whether the latent categories obtained by a PCFG-LA parser (Petrov et al., 2006) can be used as (syntactico-)semantic categories. Parsers based on Probabilistic Context-Free Grammar with Latent Categories (PCFG-LA parsers) learn latent categories when trained on a treebank. These latent categories are obtained by splitting the syntactic categories present in the treebank in a way which optimizes parsing accuracy. We use these categories to create additional word attributes.

Default grammar. The unannotated corpus was parsed with the Berkeley PCFG-LA parser (Petrov et al., 2006) using its default grammar. Latent categories for the corpus words were then extracted and used to build a lexicon with 'semantic' categories. However, the default English grammar of the Berkeley parser is not based on medical texts.

Specific grammar. We hypothesized that a better performance in disorder entity recognition could be achieved by using a grammar adapted to the specific domain and genre of our target corpus. We therefore specifically trained a grammar for this corpus. We assumed that the Charniak-McClosky parser (McClosky et al., 2006), which itself is self-trained on biomedical texts (McClosky and Charniak, 2008), should have a better performance on these texts than the vanilla Berkeley parser. We therefore prepared a training treebank by extracting 10% of the unannotated corpus parsed with the Charniak-McClosky parser, then used it with no manual correction to train a specific grammar for the Berkeley parser. This specific grammar was used by the Berkeley parser to parse the larger corpus.

As an example, one of the latent categories obtained by splitting the noun (NN) category in the corpus is NN-7, which contains (among other) the following words: *Albuterol, postop, Insulin, Pantoprazole, Lopressor, platelet, oxygen, Furosemide, Glyburide, Prednisone, Atenolol, ejection, Levofloxacin, Toprol*, most of which are drug names.

Building a lexicon of latent categories is slightly more complex than for Brown clusters. This is because depending on the context in which it occurs, a word in a sentence may receive a variety of latent categories. For each word, we collected all the categories it received in the automatically parsed corpus, and ordered them by descending number of occurrences. We only kept the categories with a number of occurrences greater than a threshold T, and selected the top N categories (possibly less if some or all of them did not pass the threshold). These ordered N categories were used to add N attributes to the word representation (again, #NA was used as a null value if less than N categories remained). Based on initial experiments we kept two combinations of T and N: $(T = 50, N = 2)$ and $(T = 100, N = 1)$. Therefore, 3 attributes were created with this method.

4.4. Design of experiments

Given the above-mentioned sets of attributes, the goals of our experiments were to answer the following questions:

- What is the performance of knowledge-based semantic classes (i.e., classes based on the UMLS or on the Wikipedia) with respect to the baseline system with no semantic classes (Section 5.1.);
- In a setting where no knowledge-based semantic classes would be available, what is the performance of data-driven semantic classes (i.e., word-based Brown clusters, dependency-based Brown clusters, and latent syntactic categories); besides, do syntax-aware semantic classes bring an advantage over non-syntax-aware semantic classes (Section 5.2.);
- What is the relative performance of knowledge-based and data-driven classes (Section 5.3.).

All experiments were performed on the task of DISORDER entity detection, using ten-fold cross-validation on the CLEF eHealth 2013 training set, with unsupervised attributes computed on the unannotated MIMIC-II corpus (see Section 3.). The results displayed in the tables are average precision, recall, and F-measure over these ten folds.

5. Results and discussion

5.1. Baseline system and knowledge-based semantic classes

Table 3 shows the results obtained with a baseline system which uses no semantic classes (NOSEM), with a system which only uses one type of knowledge-based semantic class (Wikipedia or UMLS), or with a combination of these. Our baseline system emphasizes precision over recall, a characteristic often found in CRFs.
The use of only UMLS attributes comes a few F-measure points (4pt) short of the baseline system NOSEM, which shows the importance of this knowledge source. Wikipedia-contributed attributes have a good precision too, but much lower coverage than UMLS attributes: this shows that Wikipedia is a resource of good quality which is worth considering even in a specialized domain, but that a large, dedicated resource such as the UMLS is best suited to the needs of the domain. Adding Wikipedia attributes on top

Feature set	P	R	F
NOSEM	**85.31**	**65.10**	**73.85**
Wikipedia	**79.03**	24.32	37.20
UMLS	78.22	**63.31**	**69.98**
UMLS + Wikipedia	78.98	**64.50**	**71.01**
NOSEM + Wikipedia	86.58	68.02	76.18
NOSEM + UMLS	88.10	74.23	80.57
NOSEM + UMLS + Wikipedia	**88.28**	**75.05**	**81.13**

Table 3: Performance of baseline features (NOSEM), knowledge-based semantic classes, and their combination

of UMLS attributes gains another 1pt F-measure, or 0.5pt when both are added to NOSEM: Wikipedia contributes information that is not present in the UMLS.
Combining all three types of attributes boosts the F-measures of both NOSEM and knowledge-based semantic classes by about 10pt: they contribute quite different information.

5.2. Baseline system and data-driven word classes

Table 4 shows the performance obtained by the same baseline system (NOSEM), by a system which only uses one type of data-driven word classes, or with a combination of these.

Feature set					P	R	F
NOSEM	Br	Br-dep	LA	LA-spec			
NOSEM					**85.31**	**65.10**	**73.85**
			LA		72.86	59.45	65.47
				LA-spec	74.61	62.57	68.06
	Br				77.28	71.53	74.30
		Br-dep			78.65	**71.74**	**75.04**
		Br-dep	LA		80.64	71.63	75.86
	Br		LA		80.02	72.45	76.05
	Br			LA-spec	80.48	72.44	76.25
		Br-dep		LA-spec	**81.21**	**73.20**	**77.00**
NOSEM			LA		85.32	67.09	75.11
NOSEM				LA-spec	85.53	67.72	75.59
NOSEM		Br-dep			85.52	71.72	78.01
NOSEM	Br				84.06	72.78	78.01
NOSEM		Br-dep		LA-spec	85.09	72.17	78.10
NOSEM	Br		LA		84.49	72.85	78.24
NOSEM		Br-dep	LA		84.98	72.62	78.31
NOSEM	Br			LA-spec	**85.33**	**73.47**	**78.96**

Table 4: Performance of baseline features (NOSEM), data-driven semantic classes, and their combination (LA = Latent categories with default grammar, LA-spec = Latent categories with specific grammar, Br = word-based Brown clusters, Br-dep = dependency-based Brown clusters)

Data-driven word classes alone. Latent categories alone obtain a better F-measure than Wikipedia alone because their higher recall more than compensates for a lower precision. Retraining the parser on a part of the domain-specific corpus results in categories that gain 2.5pt F-measure. Brown clusters alone outperform UMLS alone and even the baseline NOSEM attributes: in our setting, distributional analysis alone performs better than a combination of lexical, morphological, syntactic and document structure features. It does so by a strong increase in recall, at the expense of a decrease in precision. This is the inverse of the preference for precision of our baseline system. Dependency-based Brown clusters also bring an improvement of 0.7pt F-measure over bigram-based Brown clusters.
PCFG-LA latent categories are obtained by splitting syntactic categories into more specific subcategories. We may thus compare them to using only part-of-speech (POS) information. A system trained with only part-of-speech information (not displayed in Table 4) obtains (P=49.58, R=22.16, F=30.63), to be compared to the rows for LA or LA-spec which obtain much higher results. We believe that the more specific categories, added to the description of each word by three categories instead of only one POS, account for this large difference.

Combination of data-driven word classes. When combined, Brown clusters and Latent categories further gain 2pt F-measure, which shows that they capture different types of information. Further adding the NOSEM attributes gains another 2pt F-measure. All in all, this falls short of NOSEM plus knowledge-based classes by 2pt F-measure. On the one hand, we can say that rich knowledge-based classes such as obtained by the UMLS (and to a lesser extent Wikipedia) still do a better job of helping the detection of disorders, both in terms of precision and recall. On the other hand, we also observe that the huge human effort invested to build these resources only gains 2pt in our setting. Finally, we note that the combination of unsupervised word classes with the baseline NOSEM features boosts recall while leaving precision untouched.

Combination of baseline and data-driven word classes. In combination with the NOSEM attributes, dependency-based Brown clusters bring a positive or negative contribution depending on the presence or absence of LA / LA-spec attributes. In contrast, the specific grammar for latent categories improves the results compared to the default grammar in most of the cases. Nevertheless, in all cases combining Brown clusters with PCFG latent categories improves over either of them.

5.3. Baseline system and all word classes

Table 5 recalls the performance obtained by the baseline system (NOSEM, here abbreviated as N). It then shows a selection of combinations of knowledge-based semantic classes and data-driven word classes. Finally, it displays a selection of combinations of NOSEM and both types of word classes. The combination of all knowledge-based classes and all data-driven classes (last row in part two of the table) reaches 80.64 F-measure, which improves over the combination of only knowledge-based classes (F=71.01) or only data-driven classes (F=77.00). It is very close to the best result of Table 3 which combined NOSEM with knowledge-based classes (F=81.13). This probably means that data-driven classes encode most of the lexical and syntactic information provided to NOSEM. The full combination of baseline attributes (NOSEM) and both types of word classes reaches F=82.39, which improves the above-mentioned best result by 1.3pt. This shows that in

Feature set							P	R	F
N	W	U	Br	Br-dep	LA	LA-spec			
N							**85.31**	**65.10**	**73.85**
	W				LA		77.35	63.26	69.60
		U			LA		83.42	72.96	77.84
		U				LA-spec	83.89	73.37	78.28
	W	U			LA		84.16	73.61	78.54
	W	U				LA-spec	**84.70**	74.01	78.99
		U	Br				81.78	76.63	79.12
	W	U	Br				81.93	76.74	79.25
	W	U	Br			LA-spec	84.29	77.20	80.59
	W	U	Br		LA		84.27	**77.31**	**80.64**
N	W					LA-spec	86.41	71.72	78.38
N	W		Br				85.11	74.08	79.21
N	W		Br			LA-spec	85.04	74.40	79.36
N		U				LA-spec	87.94	74.87	80.88
N	W	U				LA-spec	88.13	75.69	81.44
N		U		Br-dep			87.91	76.46	81.79
N		U	Br			LA-spec	87.02	77.31	81.88
N	W	U	Br		LA		87.33	77.15	81.92
N		U	Br				87.22	77.28	81.95
N	W	U		Br-dep			88.09	76.63	81.99
N	W	U	Br				87.49	77.48	82.18
N	W	U		Br-dep		LA-spec	**88.59**	76.71	82.22
N	W	U	Br			LA-spec	87.55	**77.80**	**82.39**

Table 5: Performance of the combination of knowledge-based and data-driven classes (N = baseline with no semantic classes, W=Wikipedia, U = UMLS, LA = Latent categories with default grammar, LA-spec = Latent categories with specific grammar, Br = word-based Brown clusters, Br-dep = dependency-based Brown clusters)

the present setting where we can pool two kinds of rich knowledge-based classes, the data-driven classes only bring a moderate additional improvement. In contrast, adding knowledge-based classes (UMLS + Wikipedia) to the best combination of data-driven classes with NOSEM boosts the F-measure by 3.4pt F-measure: this underlines the importance of these classes when they are available.

Ablation studies. Table 6 reproduces the best result and some of the rows of Table 5 to show ablation experiments. The most contributing set of attributes in this final configuration is the UMLS (3pt F-measure), followed by NOSEM (1.8pt) and word-based Brown clusters (1.0pt). The remaining two sets of attributes (Wikipedia and LA-spec) contribute much less to the final F-measure (0.5pt and 0.2pt). Besides ablation, we also examined the effect of the dependency-based variant of the Brown clusters: they increase the precision of the system, but decrease its recall, which leads to a slightly lower F-measure. We also confirm that using a specific grammar instead of the default grammar for the PCFG-LA parser improves the results, with a 0.5pt gain in F-measure (in fact, using the PCFG-LA word categories with the default grammar is worse than not using them at all).

Feature set							P	R	F	$-\delta$
N	W	U	Br	Br-dep	LA	LA-spec				
N	W		Br			LA-spec	85.04	74.40	79.36	3.0
	W	U	Br			LA-spec	84.29	77.20	80.59	1.8
N	W	U				LA-spec	88.13	75.69	81.44	1.0
N		U	Br			LA-spec	87.02	77.31	81.88	0.5
N	W	U	Br		LA		87.33	77.15	81.92	0.5
N	W	U	Br				87.49	77.48	82.18	0.2
N	W	U		Br-dep		LA-spec	**88.59**	76.71	82.22	0.2
N	W	U	Br			LA-spec	87.55	**77.80**	**82.39**	0.0

Table 6: Ablation studies with respect to the best configuration (δ is the loss incurred by removing one set of attributes). See Table 5 for the other abbreviations.

Sets of features	Minimum	Median	Maximum
Wikipedia	3.17	(NA)	4.80
UMLS	2.90	16.40	21.59
LA-spec	63.16	86.40	94.08
LA	83.60	91.90	95.53
Br-dep	74.50	96.59	97.65
Br	97.91	98.87	99.30

Table 7: Coverage (in %) of the corpus by knowledge-based and data-driven attributes (see Table 5 for abbreviations)

5.4. Coverage of word classes

An important factor in understanding the systems is to measure the coverage of the attributes on the corpus. By this we mean the proportion of occurrences of tokens in the corpus for which a non-null value is provided by a lexicon for a given attribute. Knowledge-based semantic classes and data-driven semantic classes are each represented by a set of attributes: 2 for Wikipedia, 11 for UMLS, 10 for Brown clusters and 3 for PCFG-LA categories. Table 7 provides coverage information for each set of attributes: the lowest, median and highest value obtained by an attribute in this set. Since Wikipedia only has two attributes, it only has two values which are displayed in the table as minimum and maximum values.

The rows in the table are sorted by maximum value in ascending order. We observe that the medical lexicon that we built from Wikipedia has the lowest coverage, which explains its low recall. The UMLS essentially contains medical terms, most of which are built around a head noun. These properties explain the moderate coverage of UMLS attributes.

In principle, data-driven classes are susceptible to obtain full coverage of the words in the annotated corpus, since they are induced from a much larger unannotated corpus of the same origin. Their actual coverage depends on the thresholds imposed in the implemented methods. Nevertheless, for Brown clusters, even with our highest threshold of 32 occurrences in the unannotated corpus, between 98% and 99% of the word occurrences in the annotated corpus are present in a cluster. Dependency-based Brown clusters have slightly lower coverage due to the use of reduced dependencies, which result for instance in the suppression of prepositions and their replacement with reduced dependency names such as *prep_of*. The thresholds imposed on

latent categories were higher (50 and 100) and may explain the slightly lower coverage of these categories. This coverage remains high, with maximums of 94-95%.

In summary, data-driven word classes have a higher coverage of the corpus than knowledge-based semantic classes. Their contribution to the performance of the system, when added to knowledge-based classes, acts through an increase in recall. But because they are noisier, they also decrease precision. Since the system based on NOSEM plus knowledge-based classes lacks recall, they improve its F-measure.

5.5. Limitations

This study has the following limitations. First, it considers only one corpus in one domain. Similar studies should be performed on a variety of corpora to check the generalization of these observations. Second, this dataset deals with only one type of entity (disorders). The impact of the various types of word classes that we studied in this paper could also vary depending on the number of entity types present in an annotated corpus and their nature.

Besides, concerning syntax-informed word classes, i.e., dependency-based Brown clusters and PCFG-LA word categories, we note that syntactic analysis was performed on texts where sentence splitting was often incorrect, leading to some erroneous parses. Improved sentence splitting might lead to a more accurate comparison of methods. We expect however that the induced changes should remain marginal.

Finally, the ablation studies in Table 6 show that although the UMLS has a much smaller coverage of the full annotated corpus, it does provide the largest contribution to recall (the first row of the table, which shows results without UMLS, has the lowest recall), followed by Brown clusters (third row in the table). Measuring coverage over the target entities and their vicinity, instead of over the full corpus, might therefore be more relevant.

6. Conclusion

We have performed systematic experiments to study the contribution of both knowledge-based and unsupervised word classes to disorder recognition, using the CLEF eHealth 2013 challenge data. They showed that knowledge-based word classes obtained from the UMLS and Wikipedia drastically boost (+7.3pt F-measure) a baseline system which has no semantic classes (NOSEM), with a main contribution of the UMLS (6.7pt). We also observed that in the absence of knowledge-based classes, unsupervised word classes also improve the NOSEM system, albeit by a smaller margin (5.1pt), with a main contribution by Brown clusters.

We proposed two novel methods to compute word classes based on syntactic information: dependency-based Brown clusters and PCFG-LA word categories. Our experiments with dependency-based Brown clusters were not conclusive. We also observed that PCFG-LA categories brought improvements when based on a grammar that has been trained on a similar corpus. When added to the knowledge-based classes, unsupervised word classes brought a moderate additional improvement (1.3pt), highlighting the importance of the knowledge-based classes in our context.

This study should be extended to new data sets with multiple and different types of entities, and other domains or specialties. Other methods could also be tested to compute word classes based on syntactic representations, such as modifying Brown's algorithm to model syntactic dependencies in a more principled way. Neural network 'word embeddings' (e.g., (Collobert and Weston, 2008; Mnih and Hinton, 2009; Mikolov et al., 2013)) are also a popular way to derive data-driven word representations and word classes which add yet another class of methods to try to match the power of knowledge-based semantic classes.

7. References

Bodenreider, O. (2004). The Unified Medical Language System (UMLS): Integrating biomedical terminology. *Nucleic Acids Research*, 32(Database issue):D267–270.

Bodnari, A., Deléger, L., Lavergne, T., Névéol, A., and Zweigenbaum, P. (2013). A supervised named-entity extraction system for medical text. In *Proc of CLEF*.

Brown, P. F., Della Pietra, V. J., de Souza, P. V., Lai, J. C., and Mercer, R. L. (1992). Class-based n-gram models of natural language. *Computational Linguistics*, 18(4):467–79.

Collobert, R. and Weston, J. (2008). A unified architecture for natural language processing: deep neural networks with multitask learning. In *Proc of ICML*, pages 160–7.

de Bruijn, B., Cherry, C., Kiritchenko, S., Martin, J., and Zhu, X. (2011). Machine-learned solutions for three stages of clinical information extraction: the state of the art at i2b2 2010. *J Am Med Inform Assoc*, 18(5):557–562, September. Epub 2011 May 12.

De Marneffe, M.-C. and Manning, C. D., (2008). *Stanford typed dependencies manual*.

Jonnalagadda, S., Cohen, T., Wu, S., and Gonzalez, G. (2012). Enhancing clinical concept extraction with distributional semantics. *J Biomed Inform*, 45(1):129–140, February. Published online Nov 7, 2011.

Jonnalagadda, S., Cohen, T., Wu, S., Liu, H., and Gonzalez, G. (2013). Using empirically constructed lexical resources for named entity recognition. *Biomed Inform Insights*, 6(Suppl 1):17–27.

Lafferty, J. D., McCallum, A., and Pereira, F. C. N. (2001). Conditional Random Fields: Probabilistic models for segmenting and labeling sequence data. In *Proc of ICML*, pages 282–9, Williamstown, MA.

Lavergne, T., Cappé, O., and Yvon, F. (2010). Practical very large scale CRFs. In *Proc of ACL*, pages 504–13, July.

Liang, P. (2005). Semi-supervised learning for natural language. Master's thesis, MIT.

McClosky, D. and Charniak, E. (2008). Self-training for biomedical parsing. In *Proc of ACL*, pages 101–4.

McClosky, D., Charniak, E., and Johnson, M. (2006). Reranking and self-training for parser adaptation. In *Proc of Coling*.

Meystre, S. M., Savova, G. K., Kipper-Schuler, K., and Hurdle, J. (2008). Extracting information from textual

documents in the electronic health record: a review of recent research. In *Yearbook of Medical Informatics*, pages 128–144. Shattauer, Stuttgart.

Mikolov, T., Yih, W., and Zweig, G. (2013). Linguistic regularities in continuous space word representations. In Vanderwende, L., III, H. D., and Kirchhoff, K., editors, *HLT-NAACL*, pages 746–751. The Association for Computational Linguistics.

Miller, S., Guinness, J., and Zamanian, A. (2004). Name tagging with word clusters and discriminative training. In *Proc of HLTC*, pages 337–42.

Mnih, A. and Hinton, G. E. (2009). A scalable hierarchical distributed language model. *Advances in neural information processing systems*, 21:1081–88.

Petrov, S., Barrett, L., Thibaux, R., and Klein, D. (2006). Learning accurate, compact, and interpretable tree annotation. In *Proc of ACL*, pages 433–40.

Saeed, M., Villarroel, M., Reisner, A. T., Clifford, G., Lehman, L.-W. W., Moody, G., Heldt, T., Kyaw, T. H., Moody, B., , and Mark, R. G. (2011). Multiparameter Intelligent Monitoring in Intensive Care II: a public-access Intensive Care Unit database. *Critical Care Medicine*, 39(5):952–60.

Savova, G. K., Masanz, J. J., Ogren, P. V., Zheng, J., Sohn, S., Kipper-Schuler, K. C., and Chute, C. G. (2010). Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *J Am Med Inform Assoc*, 17(5):507–13.

Suominen, H., Salanterä, S., Chapman, W. W., Velupillai, S., Savova, G., Elhadad, N., Pradhan, S., South, B. R., Mowery, D. L., Jones, G. J., Leveling, J., Kelly, L., Goeuriot, L., Martinez, D., and Zuccon, G. (2013). Overview of the ShARe/CLEF eHealth evaluation lab 2013. In *Proc of CLEF*.

Sutton, C. and McCallum, A. (2006). An introduction to Conditional Random Fields for relational learning. In Getoor, L. and Taskar, B., editors, *Introduction to Statistical Relational Learning*. MIT Press.

Tang, B., Cao, H., Wu, Y., Jiang, M., and Xu, H. (2013). Recognizing clinical entities in hospital discharge summaries using Structural Support Vector Machines with word representation features. *BMC Med Inform Decis Mak*, 13(Suppl 1):S1. Published online Apr 5, 2013.

Turian, J., Ratinov, L., and Bengio, Y. (2010). Word representations: a simple and general method for semi-supervised learning. In *Proc of ACL*, pages 384–94.

Uzuner, O., South, B. R., Shen, S., and DuVall, S. L. (2011). 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *J Am Med Inform Assoc*, 18(5):552–6.

Widdows, D. and Cohen, T. (2010). The semantic vectors package: New algorithms and public tools for distributional semantics. In *Fourth IEEE International Conference on Semantic Computing (ICSC)*, pages 9–15. IEEE.