Predicting the relevance of distributional semantic similarity with contextual information

Philippe Muller
IRIT, Toulouse University
Université Paul Sabatier
118 Route de Narbonne
31062 Toulouse Cedex 04
philippe.muller@irit.fr

Cécile Fabre
CLLE, Toulouse University
Université Toulouse-Le Mirail
5 alles A. Machado
31058 Toulouse Cedex
cecile.fabre@univ-tlse2.fr

Clémentine Adam
CLLE, Toulouse University
Université Toulouse-Le Mirail
5 alles A. Machado
31058 Toulouse Cedex
clementine.adam@univ-tlse2.fr

Abstract

Using distributional analysis methods to compute semantic proximity links between words has become commonplace in NLP. The resulting relations are often noisy or difficult to interpret in general. This paper focuses on the issues of evaluating a distributional resource and filtering the relations it contains, but instead of considering it in abstracto, we focus on pairs of words in context. In a discourse, we are interested in knowing if the semantic link between two items is a by-product of textual coherence or is irrelevant. We first set up a human annotation of semantic links with or without contextual information to show the importance of the textual context in evaluating the relevance of semantic similarity, and to assess the prevalence of actual semantic relations between word tokens. We then built an experiment to automatically predict this relevance, evaluated on the reliable reference data set which was the outcome of the first annotation. We show that in-document information greatly improve the prediction made by the similarity level alone.

1 Introduction

The goal of the work presented in this paper is to improve distributional thesauri, and to help evaluate the content of such resources. A distributional thesaurus is a lexical network that lists semantic neighbours, computed from a corpus and a similarity measure between lexical items, which generally captures the similarity of contexts in which the items occur. This way of building a semantic network has been very popular since (Grefenstette, 1994; Lin, 1998), even though the nature of the information it contains is hard to define, and its evaluation is far from obvious. A distributional thesaurus includes a lot of "noise" from a semantic point of view, but also lists relevant lexical pairs that escape classical lexical relations such as synonymy or hypernymy.

There is a classical dichotomy when evaluating NLP components between extrinsic and intrinsic evaluations (Jones, 1994), and this applies to distributional thesauri (Curran, 2004; Poibeau and Messiant, 2008). Extrinsic evaluations measure the capacity of a system in which a resource or a component to evaluate has been used, for instance in this case information retrieval (van der Plas, 2008) or word sense disambiguation (Weeds and Weir, 2005). Intrinsic evaluations try to measure the resource itself with respect to some human standard or judgment, for instance by comparing a distributional resource with respect to an existing synonym dictionary or similarity judgment produced by human subjects (Pado and Lapata, 2007; Baroni and Lenci, 2010). The shortcomings of these methods have been underlined in (Baroni and Lenci, 2011). Lexical resources designed for other objectives put the spotlight on specific areas of the distributional thesaurus. They are not suitable for the evaluation of the whole range of semantic relatedness that is exhibited by distributional similarities, which exceeds the limits of classical lexical relations, even though researchers have tried to collect equivalent resources manually, to be used as a gold standard (Weeds, 2003; Bordag, 2008; Anguiano et al., 2011). One advantage of distributional similarities is to exhibit a lot of different semantic relations, not necessarily standard lexical relations. Even with respect to established lexical resources, distributional approaches may improve coverage, complicating the evaluation even more.

The method we propose here has been designed as an intrinsic evaluation with a view to validate semantic proximity links in a broad per-

Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, pages 479–488,
Baltimore, Maryland, USA, June 23-25 2014. ©2014 Association for Computational Linguistics

spective, to cover what (Morris and Hirst, 2004) call "non classical lexical semantic relations". For instance, agentive relations (author/publish, author/publication) or associative relations (actor/cinema) should be considered. At the same time, we want to filter associations that can be considered as accidental in a semantic perspective (e.g. flag and composer are similar because they appear a lot with nationality names). We do this by judging the relevance of a lexical relation in a context where both elements of a lexical pair occur. We show not only that this improves the reliability of human judgments, but also that it gives a framework where this relevance can be predicted automatically. We hypothetize that evaluating and filtering semantic relations in texts where lexical items occur would help tasks that naturally make use of semantic similarity relations, but assessing this goes beyond the present work.

In the rest of this paper, we describe the resource we used as a case study, and the data we collected to evaluate its content (section 2). We present the experiments we set up to automatically filter semantic relations in context, with various groups of features that take into account information from the corpus used to build the thesaurus and contextual information related to occurrences of semantic neighbours 3). Finally we discuss some related work on the evaluation and improvement of distributional resources (section 4).

2 Evaluation of lexical similarity in context

2.1 Data

We use a distributional resource for French, built on a 200M word corpus extracted from the French Wikipedia, following principles laid out in (Bourigault, 2002) from a structured model (Baroni and Lenci, 2010), i.e. using syntactic contexts. In this approach, contexts are triples (governor,relation,dependent) derived from syntactic dependency structures. Governors and dependents are verbs, adjectives and nouns. Multiword units are available, but they form a very small subset of the resulting neighbours. Base elements in the thesaurus are of two types: arguments (dependents' lemma) and predicates (governor+relation). This is to keep the predicate/argument distinction since similarities will be computed between predicate pairs or argument pairs, and a lexical item can appear in many predicates and as an argument (e.g. *interest* as argument, *interest_for* as one predicate). The similarity of distributions was computed with Lin's score (Lin, 1998).

We will talk of lexical neighbours or distributional neighbours to label pairs of predicates or arguments, and in the rest of the paper we consider only lexical pairs with a Lin score of at least 0.1, which means about 1.4M pairs. This somewhat arbitrary level is an *a priori* threshold to limit the resulting database, and it is conservative enough not to exclude potential interesting relations. The distribution of scores is given figure 1; 97% of the selected pairs have a score between 0.1 and 0.29.

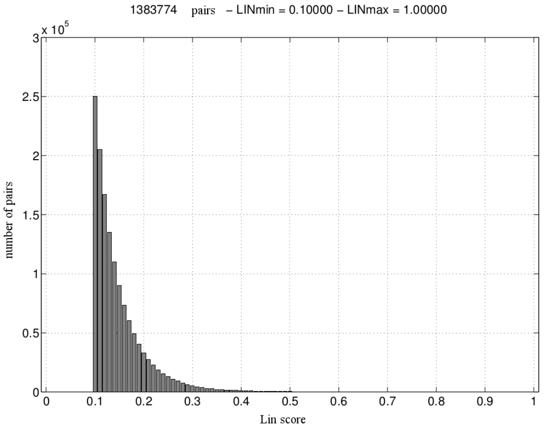


Figure 1: Histogram of Lin scores for pairs considered.

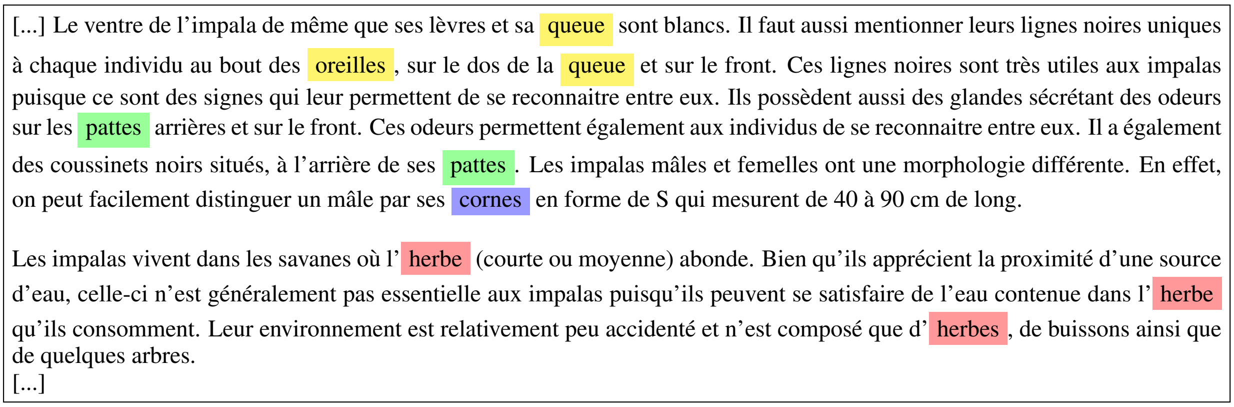
To ease the use of lexical neighbours in our experiments, we merged together predicates that include the same lexical unit, *a posteriori*. Thus there is no need for a syntactic analysis of the context considered when exploiting the resource, and sparsity is less of an issue[1].

2.2 Annotation

In order to evaluate the resource, we set up an annotation in context: pairs of lexical items are to be judged in their context of use, in texts where they occur together. To verify that this methodology is useful, we did a preliminary annotation to contrast judgment on lexical pairs with or without this contextual information. Then we made a larger annotation in context once we were assured of the reliability of the methodology.

For the preliminary test, we asked three annotators to judge the similarity of pairs of lexical items without any context (no-context), and to judge the

[1] Whenever two predicates with the same lemma have common neighbours, we average the score of the pairs.



[...] Le ventre de l'impala de même que ses lèvres et sa queue sont blancs. Il faut aussi mentionner leurs lignes noires uniques à chaque individu au bout des oreilles , sur le dos de la queue et sur le front. Ces lignes noires sont très utiles aux impalas puisque ce sont des signes qui leur permettent de se reconnaitre entre eux. Ils possèdent aussi des glandes sécrétant des odeurs sur les pattes arrières et sur le front. Ces odeurs permettent également aux individus de se reconnaitre entre eux. Il a également des coussinets noirs situés, à l'arrière de ses pattes . Les impalas mâles et femelles ont une morphologie différente. En effet, on peut facilement distinguer un mâle par ses cornes en forme de S qui mesurent de 40 à 90 cm de long.

Les impalas vivent dans les savanes où l' herbe (courte ou moyenne) abonde. Bien qu'ils apprécient la proximité d'une source d'eau, celle-ci n'est généralement pas essentielle aux impalas puisqu'ils peuvent se satisfaire de l'eau contenue dans l' herbe qu'ils consomment. Leur environnement est relativement peu accidenté et n'est composé que d' herbes , de buissons ainsi que de quelques arbres.
[...]

Figure 2: Example excerpt during the annotation of lexical pairs: annotators focus on a target item (here *corne*, horn, in blue) and must judge yellow words (pending: *oreille/queue*, ear/tail), either validating their relevance (green words: *pattes*, legs) or rejecting them (red words: *herbe*, grass). The text describes the morphology of the impala, and its habitat.

similarity of pairs presented within a paragraph where they both occur (in context). The three annotators were linguists, and two of them (1 and 3) knew about the resource and how it was built. For each annotation, 100 pairs were randomly selected, with the following constraints:

- for the no-context annotation, candidate pairs had a Lin score above 0.2, which placed them in the top 14% of lexical neighbours with respect to the similarity level.
- for the in context annotation, the only constraint was that the pairs occur in the same paragraph somewhere in the corpus used to build the resource. The example paragraph was chosen at random.

The guidelines given in both cases were the same: "Do you think the two words are semantically close ? In other words, is there a semantic relation between them, either classical (synonymy, hypernymy, co-hyponymy, meronymy, co-meronymy) or not (the relation can be paraphrased but does not belong to the previous cases) ?"

For the pre-test, agreement was rather moderate without context (the average of pairwise kappas was .46), and much better with a context (average = .68), with agreement rates above 90%. This seems to validate the feasability of a reliable annotation of relatedness in context, so we went on for a larger annotation with two of the previous annotators.

For the larger annotation, the protocol was slightly changed: two annotators were given 42 full texts from the original corpus where lexical neighbours occurred. They were asked to judge the relation between two items types, regardless of the number of occurrences in the text. This time there was no filtering of the lexical pairs beyond the 0.1 threshold of the original resource. We followed the well-known postulate (Gale et al., 1992) that all occurrences of a word in the same discourse tend to have the same sense ("one sense per discourse"), in order to decrease the annotator workload. We also assumed that the relation between these items remain stable within the document, an arguably strong hypothesis that needed to be checked against inter-annotator agreement before beginning the final annotation . It turns out that the kappa score (0.80) shows a better inter-annotator agreement than during the preliminary test, which can be explained by the larger context given to the annotator (the whole text), and thus more occurrences of each element in the pair to judge, and also because the annotators were more experienced after the preliminary test. Agreement measures are summed-up table 1. An excerpt of an example text, as it was presented to the annotators, is shown figure 2.

Overall, it took only a few days to annotate 9885 pairs of lexical items. Among the pairs that were presented to the annotators, about 11% were judged as relevant by the annotators. It is not easy to decide if the non-relevant pairs are just noise, or context-dependent associations that were not present in the actual text considered (for polysemy reasons for instance), or just low-level associations. An important aspect is thus to guarantee that there is a correlation between the sim-

Annotators	Non-contextual		Contextual	
	Agreement rate	Kappa	Agreement rate	Kappa
N1+N2	77%	0.52	91%	0.66
N1+N3	70%	0.36	92%	0.69
N2+N3	79%	0.50	92%	0.69
Average	75, 3%	0,46	91, 7%	0,68
Experts	NA	NA	90.8%	0.80

Table 1: Inter-annotator agreements with Cohen's Kappa for contextual and non-contextual annotations. N1, N2, N3 were annotators during the pre-test; expert annotation was made on a different dataset from the same corpus, only with the full discourse context.

ilarity score (Lin's score here), and the evaluated relevance of the neighbour pairs. Pearson correlation factor shows that Lin score is indeed significantly correlated to the annotated relevance of lexical pairs, albeit not strongly ($r = 0.159$).

The produced annotation[2] can be used as a reference to explore various aspects of distributional resources, with the caveat that it is as such a bit dependent on the particular resource used. We nonetheless assume that some of the relevant pairs would appear in other thesauri, or would be of interest in an evaluation of another resource.

The first thing we can analyse from the annotated data is the impact of a threshold on Lin's score to select relevant lexical pairs. The resource itself is built by choosing a cut-off which is supposed to keep pairs with a satisfactory similarity, but this threshold is rather arbitrary. Figure 3 shows the influence of the threshold value to select relevant pairs, when considering precision and recall of the pairs that are kept when choosing the threshold, evaluated against the human annotation of relevance in context. In case one wants to optimize the F-score (the harmonic mean of precision and recall) when extracting relevant pairs, we can see that the optimal point is at .24 for a threshold of .22 on Lin's score. This can be considered as a baseline for extraction of relevant lexical pairs, to which we turn in the following section.

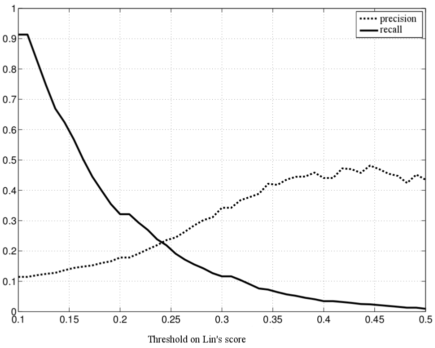


Figure 3: Precision and recall on relevant links with respect to a threshold on the similarity measure (Lin's score)

3 Experiments: predicting relevance in context

The outcome of the contextual annotation presented above is a rather sizeable dataset of validated semantic links, and we showed these linguistic judgments to be reliable. We used this dataset to set up a supervised classification experiment in order to automatically predict the relevance of a semantic link in a given discourse. We present now the list of features that were used for the model. They can be divided in three groups, according to their origin: they are computed from the whole corpus, gathered from the distributional resource, or extracted from the considered text which contains the semantic pair to be evaluated.

3.1 Features

For each pair *neighbour$_a$/neighbour$_b$*, we computed a set of features from Wikipedia (the corpus used to derive the distributional similarity): We first computed the frequencies of each item in the corpus, $freq_a$ and $freq_b$, from which we derive

- $freq_{min}$, $freq_{max}$: the min and max of $freq_a$ and $freq_b$;
- $freq_\times$: the combination of the two, or $\log(freq_a \times freq_b)$

[2] Freely available here `http://www.irit.fr/~Philippe.Muller/resources.html`.

We also measured the syntagmatic association of *neighbour*$_a$ and *neighbour*$_b$, with a mutual information measure (Church and Hanks, 1990), computed from the cooccurrence of two tokens within the same paragraph in Wikipedia. This is a rather large window, and thus gives a good coverage with respect to the neighbour database (70% of all pairs).

A straightforward parameter to include to predict the relevance of a link is of course the similarity measure itself, here Lin's information measure. But this can be complemented by additional information on the similarity of the neighbours, namely:

- each neighbour productivity : $prod_a$ and $prod_b$ are defined as the numbers of neighbours of respectively *neighbour*$_a$ and *neighbour*$_b$ in the database (thus related tokens with a similarity above the threshold), from which we derive three features as for frequencies: the min, the max, and the log of the product. The idea is that neighbours whith very high productivity give rise to less reliable relations.

- the ranks of tokens in other related items neighbours: $rank_{a-b}$ is defined as the rank of *neighbour*$_a$ among neighbours of *neighbour*$_b$ ordered with respect to Lin's score; $rang_{b-a}$ is defined similarly and again we consider as features the min, max and log-product of these ranks.

We add two categorial features, of a more linguistic nature:

- $cats$ is the pair of part-of-speech for the related items, e.g. to distinguish the relevance of NN or VV pairs.

- $predarg$ is related to the predicate/argument distinction: are the related items predicates or arguments ?

The last set of features derive from the occurrences of related tokens in the considered discourses:

First, we take into account the frequencies of items within the text, with three features as before: the min of the frequencies of the two related items, the max, and the log-product. Then we consider a *tf·idf* (Salton et al., 1975) measure, to evaluate the specificity and arguably the importance of a word

Feature	Description
$freq_{\min}$	$\min(freq_a, freq_b)$
$freq_{\max}$	$\max(freq_a, freq_b)$
$freq_{\times}$	$\log(freq_a \times freq_b)$
im	$im = \log \frac{P(a,b)}{P(a) \cdot P(b)}$
lin	Lin's score
$rank_{\min}$	$\min(rank_{a-b}, rank_{b-a})$
$rank_{\max}$	$\max(rank_{a-b}, rank_{b-a})$
$rank_{\times}$	$\log(rank_{a-b} \times rank_{b-a})$
$prod_{\min}$	$\min(prod_a, prod_b)$
$prod_{\max}$	$\max(prod_a, prod_b)$
$prod_{\times}$	$\log(prod_a \times prod_b)$
$cats$	neighbour pos pair
$predarg$	predicate or argument
$freqtxt_{\min}$	$\min(freqtxt_a, freqtxt_b)$
$freqtxt_{\max}$	$\max(freqtxt_a, freqtxt_b)$
$freqtxt_{\times}$	$\log(freqtxt_a \times freqstxt_b)$
tf·ipf	*tf·ipf*(*neighbour*$_a$)×*tf·ipf*(*neighbour*$_b$)
copr$_{ph}$	copresence in a sentence
copr$_{para}$	copresence in a paragraph
sd	smallest distance between *neighbour*$_a$ and *neighbour*$_b$
gd	highest distance between *neighbour*$_a$ and *neighbour*$_b$
ad	average distance between *neighbour*$_a$ and *neighbour*$_b$
$prodtxt_{\min}$	$\min(prod_a, prod_b)$
$prodtxt_{\max}$	$\max(prod_a, prod_b)$
$prodtxt_{\times}$	$\log(prod_a \times prod_b)$
cc	belong to the same lexical connected component

Table 2: Summary of features used in the supervised model, with respect to two lexical items a and b. The first group is corpus related, the second group is related to the distributional database, the third group is related to the textual context. Freq is related to the frequencies in the corpus, Freqtext the frequencies in the considered text.

in a document or within a document. Several variants of *tf·idf* have been proposed to adapt the measure to more local areas in a text with respect to the whole document. For instance (Dias et al., 2007) propose a *tf·isf*(*term frequency · inverse sentence frequency*), for topic segmentation. We similarly defined a *tf·ipf* measure based on the frequency of a word within a paragraph with respect to its frequency within the text. The resulting feature we used is the product of this measure for *neighbour*$_a$ and *neighbour*$_b$.

A few other contextual features are included in the model: the distances between pairs of related items, instantiated as:

- distance in words between occurrences of related word types:
  - minimal distance between two occurrences (sd)
  - maximal distance between two occurrences (gd)
  - average distance (ad) ;
- boolean features indicating whether *neighbour*$_a$ and *neighbour*$_b$ appear in the same sentence ($copr_s$) or the same paragraph ($copr_{para}$).

Finally, we took into account the network of related lexical items, by considering the largest sets of words present in the text and connected in the database (self-connected components), by adding the following features:

- the degree of each lemma, seen as a node in this similarity graph, combined as above in minimal degree of the pair, maximal degree, and product of degrees ($prodtxt_{\min}$, $prodtxt_{\max}$, $prodtxt_{\times}$). This is the number of pairs (present in the text) where a lemma appears in.

- a boolean feature cc saying whether a lexical pair belongs to a connected component of the text, except the largest. This reflects the fact that a small component may concern a lexical field which is more specific and thus more relevant to the text.

  Figure 4 shows examples of self-connected components in an excerpt of the page on *Gorille* (gorilla), e.g. the set {*pelage, dos, fourrure*} (coat, back, fur).

The last feature is probably not entirely independent from the productivity of an item, or from the tf.ipf measure.

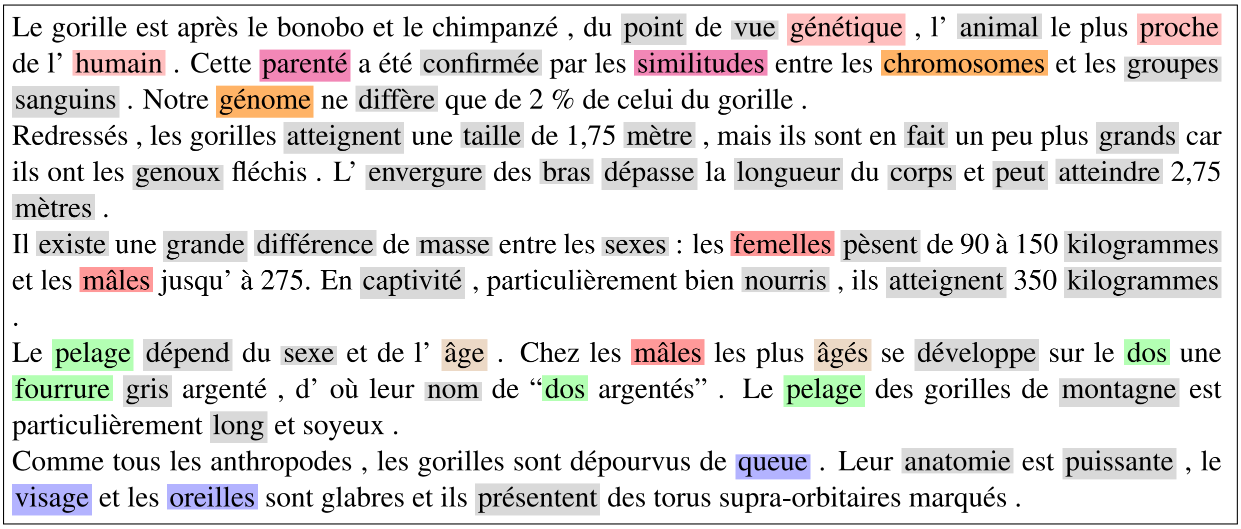
Table 2 sums up the features used in our model.

3.2 Model

Our task is to identify relevant similarities between lexical items, between all possible related pairs, and we want to train an inductive model, a classifier, to extract the relevant links. We have seen that the relevant/not relevant classification is very imbalanced, biased towards the “not relevant” category (about 11%/89%), so we applied methods dedicated to counter-balance this, and will focus on the precision and recall of the predicted relevant links.

Following a classical methodology, we made a 10-fold cross-validation to evaluate robustly the performance of the classifiers. We tested a few popular machine learning methods, and report on two of them, a naive bayes model and the best method on our dataset, the Random Forest classifier (Breiman, 2001). Other popular methods (maximum entropy, SVM) have shown slightly inferior combined F-score, even though precision and recall might yield more important variations. As a baseline, we can also consider a simple threshold on the lexical similarity score, in our case Lin’s measure, which we have shown to yield the best F-score of 24% when set at 0.22.

To address class imbalance, two broad types of methods can be applied to help the model focus on the minority class. The first one is to resample the training data to balance the two classes, the second one is to penalize differently the two classes during training when the model makes a mistake (a mistake on the minority class being made more costly than on the majority class). We tested the two strategies, by applying the classical Smote method of (Chawla et al., 2002) as a kind of resampling, and the ensemble method MetaCost of (Domingos, 1999) as a cost-aware learning method. Smote synthetizes and adds new instances similar to the minority class instances and is more efficient than a mere resampling. MetaCost is an interesting meta-learner that can use any classifier as a base classifier. We used Weka’s implementations of these methods (Frank et al., 2004), and our experiments and comparisons are thus easily replicated on our dataset, provided with this paper, even though they can be improved by



Le gorille est après le bonobo et le chimpanzé , du point de vue génétique , l' animal le plus proche de l' humain . Cette parenté a été confirmée par les similitudes entre les chromosomes et les groupes sanguins . Notre génome ne diffère que de 2 % de celui du gorille .
Redressés , les gorilles atteignent une taille de 1,75 mètre , mais ils sont en fait un peu plus grands car ils ont les genoux fléchis . L' envergure des bras dépasse la longueur du corps et peut atteindre 2,75 mètres .
Il existe une grande différence de masse entre les sexes : les femelles pèsent de 90 à 150 kilogrammes et les mâles jusqu' à 275. En captivité , particulièrement bien nourris , ils atteignent 350 kilogrammes .
Le pelage dépend du sexe et de l' âge . Chez les mâles les plus âgés se développe sur le dos une fourrure gris argenté , d' où leur nom de "dos argentés" . Le pelage des gorilles de montagne est particulièrement long et soyeux .
Comme tous les anthropodes , les gorilles sont dépourvus de queue . Leur anatomie est puissante , le visage et les oreilles sont glabres et ils présentent des torus supra-orbitaires marqués .

Figure 4: A few connected lexical components of the similarity graph, projected on a text, each in a different color. The groups are, in order of appearance of the first element: {genetic, close, human}, {similarity, kinship}, {chromosome, genome}, {male, female}, {coat, back, fur}, {age/N, aged/A}, {ear, tail, face}. The text describes the gorilla species, more particularly its morphology. Gray words are other lexical elements in the neighbour database.

refinements of these techniques. We chose the following settings for the different models: naive bayes uses a kernel density estimation for numerical features, as this generally improves performance. For Random Forests, we chose to have ten trees, and each decision is taken on a randomly chosen set of five features. For resampling, Smote advises to double the number of instances of the minority class, and we observed that a bigger resampling degrades performances. For cost-aware learning, a sensible choice is to invert the class ratio for the cost ratio, i.e. here the cost of a mistake on a relevant link (false negative) is exactly 8.5 times higher than the cost on a non-relevant link (false positive), as non-relevant instances are 8.5 times more present than relevant ones.

3.3 Results

We are interested in the precision and recall for the "relevant" class. If we take the best simple classifier (random forests), the precision and recall are 68.1% and 24.2% for an F-score of 35.7%, and this is significantly beaten by the Naive Bayes method as precision and recall are more even (F-score of 41.5%). This is already a big improvement on the use of the similarity measure alone (24%). Also note that predicting every link as relevant would result in a 2.6% precision, and thus a 5% F-score. The random forest model is significantly improved by the balancing techniques: the overall best F-score of 46.3% is reached with Random Forests and the cost-aware learning method. Table 3 sums up the scores for the different configurations, with precision, recall, F-score and the confidence interval on the F-score. We analysed the learning curve by doing a cross-validation on reduced set of instances (from 10% to 90%); F1-scores range from 37.3% with 10% of instances and stabilize at 80%, with small increment in every case.

The filtering approach we propose seems to yield good results, by augmenting the similarity built on the whole corpus with signals from the local contexts and documents where related lexical items appear together.

To try to analyse the role of each set of features, we repeated the experiment but changed the set of features used during training, and results are shown table 4 for the best method (RF with cost-aware learning).

We can see that similarity-related features (measures, ranks) have the biggest impact, but the other ones also seem to play a significant role. We can draw the tentative conclusion that the quality of distributional relations depends on the contextualizing of the related lexical items, beyond just the similarity score and the ranks of items as neighbours of other items.

Method	Precision	Recall	F-score	CI
Baseline (Lin threshold)	24.0	24.0	24.0	
RF	**68.1**	24.2	35.7	± 3.4
NB	34.8	51.3	41.5	± 2.6
RF+resampling	56.6	32.0	40.9	± 3.3
NB+resampling	32.8	54.0	40.7	± 2.5
RF+cost aware learning	40.4	54.3	**46.3**	± 2.7
NB+cost aware learning	27.3	**61.5**	37.8	± 2.2

Table 3: Classification scores (%) on the relevant class. CI is the confidence interval on the F-score (RF = Random Forest, NB= naive bayes).

Features	Prec.	Recall	F-score
all	40.4	54.3	46.3
all − corpus feat.	37.4	52.8	43.8
all − similarity feat.	36.1	49.5	41.8
all − contextual feat.	36.5	54.8	43.8

Table 4: Impact of each group of features on the best scores (%) : the lowest the results, the bigger the impact of the removed group of features.

4 Related work

Our work is related to two issues: evaluating distributional resources, and improving them. Evaluating distributional resources is the subject of a lot of methodological reflection (Sahlgren, 2006), and as we said in the introduction, evaluations can be divided between extrinsic and intrinsic evaluations. In extrinsic evaluations, models are evaluated against benchmarks focusing on a single task or a single aspect of a resource: either discriminative, TOEFL-like tests (Freitag et al., 2005), analogy production (Turney, 2008), or synonym selection (Weeds, 2003; Anguiano et al., 2011; Ferret, 2013; Curran and Moens, 2002). In intrinsic evaluations, associations norms are used, such as the 353 word-similarity dataset (Finkelstein et al., 2002), e.g. (Pado and Lapata, 2007; Agirre et al., 2009), or specifically designed test cases, as in (Baroni and Lenci, 2011). We differ from all these evaluation procedures as we do not focus on an *essential* view of the relatedness of two lexical items, but evaluate the link in a context where the relevance of the link is in question, an “existential” view of semantic relatedness.

As for improving distributional thesauri, outside of numerous alternate approaches to the construction, there is a body of work focusing on improving an existing resource, for instance reweighting context features once an initial thesaurus is built (Zhitomirsky-Geffet and Dagan, 2009), or post-processing the resource to filter bad neighbours or re-ranking neighbours of a given target (Ferret, 2013). They still use “essential” evaluation measures (mostly synonym extraction), although the latter comes close to our work since it also trains a model to detect (intrinsically) bad neighbours by using example sentences with the words to discriminate. We are not aware of any work that would try to evaluate differently semantic neighbours according to the context they appear in.

5 Conclusion

We proposed a method to reliably evaluate distributional semantic similarity in a broad sense by considering the validation of lexical pairs in contexts where they both appear. This helps cover non classical semantic relations which are hard to evaluate with classical resources. We also presented a supervised learning model which combines global features from the corpus used to built a distributional thesaurus and local features from the text where similarities are to be judged as relevant or not to the coherence of a document. It seems from these experiments that the quality of distributional relations depends on the contextualizing of the related lexical items, beyond just the similarity score and the ranks of items as neighbours of other items. This can hopefully help filter out lexical pairs when word lexical similarity is used as an information source where context is important: lexical disambiguation (Miller et al., 2012), topic segmentation (Guinaudeau et al., 2012). This can also be a preprocessing step when looking for similarities at higher levels, for instance at the sentence level (Mihalcea et al., 2006) or other macro-textual level (Agirre et al., 2013), since these are always aggregation functions of word similarities. There are limits to what is presented here: we need to evaluate the importance of the level of noise in the distributional neighbours database, or at least the quantity of non-semantic relations present, and this depends on the way the database is built. Our starting corpus is relatively small compared to current efforts in this framework. We are confident that the same methodology can be followed, even though the quantitative results may vary, since it is independent of the particular distributional thesaurus we used, and the way the similarities are computed.

References

E. Agirre, E. Alfonseca, K. Hall, J. Kravalova, M. Paşca, and A. Soroa. 2009. A study on similarity and relatedness using distributional and wordnet-based approaches. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 19–27. Association for Computational Linguistics.

Eneko Agirre, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, and Weiwei Guo. 2013. *sem 2013 shared task: Semantic textual similarity. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*, pages 32–43, Atlanta, Georgia, USA, June. Association for Computational Linguistics.

E.H. Anguiano, P. Denis, et al. 2011. FreDist: Automatic construction of distributional thesauri for French. In *Actes de la 18eme conférence sur le traitement automatique des langues naturelles*, pages 119–124.

M. Baroni and A. Lenci. 2010. Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics*, 36(4):673–721.

M. Baroni and A. Lenci. 2011. How we BLESSed distributional semantic evaluation. *GEMS 2011*, pages 1–10.

Stefan Bordag. 2008. A comparison of co-occurrence and similarity measures as simulations of context. In Alexander F. Gelbukh, editor, *CICLing*, volume 4919 of *Lecture Notes in Computer Science*, pages 52–63. Springer.

D. Bourigault. 2002. UPERY : un outil d'analyse distributionnelle tendue pour la construction d'ontologies partir de corpus. In *Actes de la 9e confrence sur le Traitement Automatique de la Langue Naturelle*, pages 75–84, Nancy.

Leo Breiman. 2001. Random forests. *Machine Learning*, 45(1):5–32.

Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. 2002. Smote: Synthetic minority over-sampling technique. *J. Artif. Intell. Res. (JAIR)*, 16:321–357.

Kenneth Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):pp. 22–29.

James R. Curran and Marc Moens. 2002. Improvements in automatic thesaurus extraction. In *Proceedings of the ACL-02 Workshop on Unsupervised Lexical Acquisition*, pages 59–66.

J.R. Curran. 2004. *From distributional to semantic similarity*. Ph.D. thesis, University of Edinburgh.

Gaël Dias, Elsa Alves, and José Gabriel Pereira Lopes. 2007. Topic segmentation algorithms for text summarization and passage retrieval: an exhaustive evaluation. In *Proceedings of the 22nd national conference on Artificial intelligence - Volume 2*, AAAI'07, pages 1334–1339. AAAI Press.

Pedro Domingos. 1999. Metacost: A general method for making classifiers cost-sensitive. In Usama M. Fayyad, Surajit Chaudhuri, and David Madigan, editors, *KDD*, pages 155–164. ACM.

Olivier Ferret. 2013. Identifying bad semantic neighbors for improving distributional thesauri. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 561–571, Sofia, Bulgaria, August. Association for Computational Linguistics.

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. 2002. Placing search in context: the concept revisited. *ACM Trans. Inf. Syst.*, 20(1):116–131.

Eibe Frank, Mark Hall, , and Len Trigg. 2004. Weka 3.3: Data mining software in java. www.cs.waikato.ac.nz/ml/weka/.

Dayne Freitag, Matthias Blume, John Byrnes, Edmond Chow, Sadik Kapadia, Richard Rohwer, and Zhiqiang Wang. 2005. New experiments in distributional representations of synonymy. In *Proceedings of CoNLL*, pages 25–32, Ann Arbor, Michigan, June. Association for Computational Linguistics.

W. Gale, K. Church, and D. Yarowsky. 1992. One sense per discourse. In *In Proceedings of the 4th DARPA Speech and Natural Language Workshop, New-York*, pages 233–237.

G. Grefenstette. 1994. *Explorations in automatic thesaurus discovery*. Kluwer Academic Pub., Boston.

Camille Guinaudeau, Guillaume Gravier, and Pascale Sébillot. 2012. Enhancing lexical cohesion measure with confidence measures, semantic relations and language model interpolation for multimedia spoken content topic segmentation. *Computer Speech & Language*, 26(2):90–104.

Karen Sparck Jones. 1994. Towards better NLP system evaluation. In *Proceedings of the Human Language Technology Conference*, pages 102–107. Association for Computational Linguistics.

D. Lin. 1998. An information-theoretic definition of similarity. In *Proceedings of the 15th International Conference on Machine Learning*, pages 296–304, Madison.

Rada Mihalcea, Courtney Corley, and Carlo Strapparava. 2006. Corpus-based and knowledge-based measures of text semantic similarity. In *Proceedings of the 21st national conference on Artificial intelligence, AAAI06*, volume 1, pages 775–780. AAAI Press.

Tristan Miller, Chris Biemann, Torsten Zesch, and Iryna Gurevych. 2012. Using distributional similarity for lexical expansion in knowledge-based word sense disambiguation. In *Proceedings of COLING 2012*, pages 1781–1796, Mumbai, India, December. The COLING 2012 Organizing Committee.

J. Morris and G. Hirst. 2004. Non-classical lexical semantic relations. In *Proceedings of the HLT Workshop on Computational Lexical Semantics*, pages 46–51, Boston.

Sebastian Pado and Mirella Lapata. 2007. Dependency-based construction of semantic space models. *Computational Linguistics*, 33(2):161–199.

Thierry Poibeau and Cédric Messiant. 2008. Do we still Need Gold Standards for Evaluation? In *Proceedings of the Language Resource and Evaluation Conference*.

Magnus Sahlgren. 2006. Towards pertinent evaluation methodologies for word-space models. In *In Proceedings of the 5th International Conference on Language Resources and Evaluation*.

G. Salton, C. S. Yang, and C. T. Yu. 1975. A theory of term importance in automatic text analysis. *Journal of the American Society for Information Science*, 26(1):33–44.

Peter D. Turney. 2008. A uniform approach to analogies, synonyms, antonyms, and associations. In *Proceedings of the 22nd International Conference on Computational Linguistics - Volume 1*, COLING '08, pages 905–912, Stroudsburg, PA, USA. Association for Computational Linguistics.

L. van der Plas. 2008. *Automatic Lexico-Semantic Acquisition for Question Answering*. Ph.D. thesis, University of Groningen.

J. Weeds and D. Weir. 2005. Co-occurrence retrieval: A flexible framework for lexical distributional similarity. *Computational Linguistics*, 31(4):439–475.

Julie Elizabeth Weeds. 2003. *Measures and Applications of Lexical Distributional Similarity*. Ph.D. thesis, University of Sussex.

Maayan Zhitomirsky-Geffet and Ido Dagan. 2009. Bootstrapping distributional feature vector quality. *Computational Linguistics*, 35(3):435–461.