Combining Different Features of Idiomaticity for the Automatic Classification of Noun+Verb Expressions in Basque

Antton Gurrutxaga
Elhuyar Foundation
Zelai Haudi 3, Osinalde industrialdea
Usurbil 20170. Basque Country
a.gurrutxaga@elhuyar.com

Iñaki Alegria
IXA group, Univ. of the Basque Country
Manuel Lardizabal 1
Donostia 20018. Basque Country
i.alegria@ehu.es

Abstract

We present an experimental study of how different features help measuring the idiomaticity of noun+verb (NV) expressions in Basque. After testing several techniques for quantifying the four basic properties of multiword expressions or MWEs (institutionalization, semantic non-compositionality, morphosyntactic fixedness and lexical fixedness), we test different combinations of them for classification into idioms and collocations, using Machine Learning (ML) and feature selection. The results show the major role of distributional similarity, which measures compositionality, in the extraction and classification of MWEs, especially, as expected, in the case of idioms. Even though cooccurrence and some aspects of morphosyntactic flexibility contribute to this task in a more limited measure, ML experiments make benefit of these sources of knowledge, allowing to improve the results obtained using exclusively distributional similarity features.

1 Introduction

Idiomaticity is considered the defining feature of the concept of multiword expressions (MWE). It is described as a non-discrete magnitude, whose "value" depends on a combination of features like institutionalization, non-compositionality and lexico-syntactic fixedness (Granger and Paquot, 2008).

Idiomaticity appears as a continuum rather than as a series of discrete values. Thus, the classification of MWEs into discrete categories is a difficult task. A very schematic classification that has achieved a fair degree of general acceptance among experts distinguishes two main types of MWEs at phrase-level: idioms and collocations.

This complexity of the concept of idiomaticity has posed a challenge to the development of methods addressing the measurement of the aforementioned four properties. Recent research has resulted in this issue nowadays being usually addressed through measuring the following phenomena: (i) cooccurrence, for institutionalization; (ii) distributional similarity, for non-compositionality; (iii) deviation from the behavior of free combinations, for morphosyntactic fixedness; and (iv) substitutability, for lexical fixedness. This is the broad context of our experimental work on the automatic classification of NV expressions in Basque.

2 Related Work

2.1 Statistical Idiosyncrasy or Institutionalization

Using the cooccurrence of the components of a combination as a heuristic of its institutionalization goes back to early research on this field (Church and Hanks, 1990), and is computed using association measures (AM), usually in combination with linguistic techniques, which allows the use of lemmatized and POS-tagged corpora, or the use of syntactic dependencies (Seretan, 2011). In recent years, the comparative analysis of AMs (Evert, 2005) and the combination of them (Lin et al., 2008; Pecina, 2010) have aroused considerable interest.

This approach has been recently explored in Basque (Gurrutxaga and Alegria, 2011).

Proceedings of the 9th Workshop on Multiword Expressions (MWE 2013), pages 116–125,
Atlanta, Georgia, 13-14 June 2013.

2.2 Compositionality

The central concept in characterizing compositionality is the hypothesis of distributional similarity (DS) As proposed by Baldwin and Kim (2010), "the underlying hypothesis is that semantically idiomatic MWEs will occur in markedly different lexical contexts to their component words."

Berry-Rogghe (1974) proposed R-value to measure the compositionality of verb-particle constructions (VPCs), by dividing the overlap between the sets of collocates associated with the particle by the total number of collocates of the VPC. Wulff (2010) proposes two extensions to the R-value in her research on verb-preposition-noun constructions, combining and weighting in different ways individual R-values of each component.

The Vector Space Model (VSM) is applied, among others, by Fazly and Stevenson (2007), who use the cosine as a similarity measure. The shared task Distributional Semantics and Compositionality (DiSCo) at ACL-HLT 2011 shows a variety of techniques for this task, mainly association measures and VSM (Biemann and Giesbrecht, 2011). LSA (Latent Semantic Analysis) is used in several studies (Baldwin et al., 2003; Katz and Giesbrecht, 2006; Schone and Jurafsky, 2001).

Those approaches have been applied recently to Basque (Gurrutxaga and Alegria, 2012)

2.3 Morphosyntactic Flexibility (MSFlex)

Morphosyntactic fixedness is usually computed in terms of relative flexibility, as the statistical distance between the behavior of the combination and (i) the average behavior of the combinations with equal POS composition (Fazly and Stevenson, 2007; Wulff, 2010), or (ii) the average behavior of the combinations containing each one of the components of the combination (Bannard, 2007).

Fazly and Stevenson (2007) use Kullback-Leibler divergence (KL-div) to compute this distance. They analyze a set of patterns: determination (*a/the*), demonstratives, possessives, singular/plural and passive. They compute two additional measurements (dominant pattern and presence of absence of adjectival modifiers preceding the noun).

Wulff (2010) considers (i) tree-syntactic, (ii) lexico-syntactic and (iii) morphological flexibilities, and implements two metrics for these features: (i) an extension of Barkema proposal (NSSD, normalized sum of squared deviations), (ii) a special conception of "relative entropy" (H_{rel}).

Bannard (2007), using CPMI (conditional pointwise mutual information), analyses these variants: (i) variation, addition or dropping of a determiner; (ii) internal modification of the noun phrase; and (iii) verb passivation.

2.4 Lexical Flexibility (LFlex)

The usual procedure for measuring lexical flexibility is to compute the substitutability of each component of the combination using as substitutes its synonymous, quasi-synonyms, related words, etc.

The pioneering work in this field is Lin (1999), who uses a thesaurus automatically built from text. This resource is used in recent research (Fazly and Stevenson, 2007). They assume that the target pair is lexically fixed to the extent that its PMI deviates from the average PMI of its variants generated by lexical substitution. They compute flexibility using the z-score.

In Van de Cruys and Moirón (2007), a technique based on KL-div is used for Duch. They define R_{nv} as the ratio of noun preference for a particular verb (its KL-div), compared to the other nouns that are present in the cluster of substitutes. Similarly for R_{vn}. The substitute candidates are obtained from the corpus using standard distributional similarity techniques.

2.5 Other Methods

Fazly and Stevenson (2007) consider two other features: (i) the verb itself; and (ii) the semantic category of the noun according to WordNet.

2.6 Combined Systems

In order to combine several sources of knowledge, several studies have experimented with using Machine Learning methods (ML).

For Czech, Pecina (2010) combines only AMs using neural networks, logistic regression and SVM (Support Vector Machine). Lin et al. (2008) employ logistic linear regression model (LLRM) to combine scores of AMs.

Venkatapathy and Joshi (2005) propose a minimally supervised classification scheme that incorporates a variety of features to group verb-noun combinations. Their features drawn from AM and DS, but some of each type are tested and combined. They compute ranking correlation using SVM, achieving results of about 0.45.

Fazly and Stevenson (2007) use all the types of knowledge, and decision trees (C5.0) as a learning method, and achieve average results (F-score) near to 0.60 for 4 classes (literal, abstract, light verbs and idioms). The authors claim that the syntactic and combined fixedness measures substantially outperform measures of collocation extraction.

3 Experimental Setup

3.1 Corpus and Preprocessing

We use a journalistic corpus of 75 million words (MW) from two sources: (1) Issues published in 2001-2002 by the newspaper *Euskaldunon Egunkaria* (28 MW); and (2) Issues published in 2006-2010 by the newspaper *Berria* (47 MW).

The corpus is annotated with lemma, POS, fine grained POS (subPOS), case and number information using Eustagger developed by the IXA group of the University of the Basque Country. A precision of 95.42% is reported for POS + subPOS + case analysis (Oronoz et al., 2010).

3.2 Extraction of Bigram Candidates

The key data for defining a Basque NV bigram are lemma and case for the noun, and lemma for the verb. Case data is needed to differentiate, for example, *kontu hartu* ("to ask for an explanation") from *kontuan hartu* ("to take into account"), where *kontu* is a noun lemma in the inessive case.

In order to propose canonical forms, we need, for nouns, token, case and number annotations in bigram data. Those canonical forms can be formulated using number normalization, as described in Gurrutxaga and Alegria (2011). Bigrams belonging to the same key noun_lemma/noun_case+verb_lemma are normalized; a single bigram with the most frequent form is created, and the frequencies of bigrams and those of the noun unigrams summed.

We use the Ngram Statistics Package-NSP (Banerjee and Pedersen, 2010) to generate NV bigrams from a corpus generated from the output of Eustagger. Taking into account our previous results (Gurrutxaga and Alegria, 2011), we use a window span of ±1 and a frequency threshold of $f > 30$. Before generation, some surface-grammar rules are applied to correct annotations that produce noise. For example, in most Basque AdjN combinations, the adjetive is a verb in a participe form (eg. *indar armatuak*, 'armed forces'). Similarly, those kind of participles can function as nouns (*gobernuaren aliatuak*, 'the allies of the government'). Not tagging those participles properly would introduce noise in the extraction of NV combinations.

3.3 Experiments Using Single Knowledge Sources

3.3.1 Cooccurrence

The cooccurrence data provided by NSP in the bigram extraction step is processed to calculate AMs. To accomplish this, we use Stefan Evert's UCS toolkit (Evert, 2005). The most common AMs are calculated: f, t-score, log-likelihood ratio, MI, MI^3, and chi-square (χ^2).

3.3.2 Distributional Similarity

The idea is to compare the contexts of each NV bigram with the contexts of its corresponding components, by means of different techniques. The more similar the contexts, the more compositional the combination.

Context Generation We extract the context words of each bigram from the sentences with contiguous cooccurrences of the components. The noun has to occur in the grammatical case in which it has been defined after bigram normalization.

The contexts of the corresponding noun and verb are extracted separately from sentences where they did not occur together. Only content-bearing lemmas are included in the contexts (nouns, verbs and adjectives).

Context Comparison We process the contexts in two different ways:

First, we construct a VSM model, representing the contexts as vectors. As similarity measures, we use Berry-Roghe's R-value (R_{BR}) and the two extensions to it proposed by Wulff (R_{W1} and R_{W2}), Jaccard index and cosine. For the cosine, different AMs have been tested for vector weights (f, t-score, LLR and PMI). We experiment with different percentages of the vector and different numbers of collocates, using the aforementioned measures to rank the collocates. The 100 most frequent words in the corpus are stopped.

Second, we represent the same contexts as documents, and compare them by means of different indexes using the Lemur Toolkit (Allan et al., 2003). The contexts of the bigrams are used as queries against a document collection containing the context-documents of all the members of the bigrams. This can be implemented in different ways; the best results were obtained using the following:

- Lemur_1 (L1): As with vectors, the contexts of a bigram are included in a single query document, and the same is done for the contexts of its members

- Lemur_2 (L2): The context sentences of bigrams are treated as individual documents, but the contexts of each one of its members are represented in two separate documents

Due to processing reasons, the number of context sentences used in Lemur to generate documents is limited to 2,000 (randomly selected from the whole set of contexts).

We further tested LSA (using Infomap[1]), but the above methods yielded better results.

3.3.3 Morphosyntactic Flexibility

We focus on the variation of the N slot, distinguishing the main type of extensions and number inflections. Among left-extensions, we take into account relative clauses. In addition, we consider the order of components as a parameter. We present some examples of the free combination *liburua irakurri* ("to read a book")

- Determiner: *liburu bat irakurri dut* ("I have read one book"), *zenbat liburu irakurri dituzu?* ("how many books have you read?")

- Postnominal adjective: *liburu interesgarria irakurri nuen* ("I read an interesting book")

- Prenominal adjective: *italierazko liburua irakurri* ("to read a book in Italian")

- Relative clause: *irakurri dudan liburua* ("the book I have read"), *anaiak irakurritako liburu batzuk* ("some books read by my brother")

- Number inflection: *liburua/liburuak/ liburu/liburuok irakurri* ("to read a/some/∅/these book(s)")

- Order of components (NV / VN): *liburua irakurri dut / irakurri dut liburua* ("I have read a book")

We count the number of variations for each bigram, for all NV bigrams, and for each combination of the type bigram_component+POS of the other component (e.g, for *liburua irakurri*, the variations of all the combinations *liburua*+V and N+*irakurri*).

To calculate flexibility, we experiment with all the measures described in section 2.3: Fazly's KL-div, Wulff's NSSD and Hrel (relative entropy), and Banard's CPMI.

3.3.4 Lexical Flexibility

In order to test the substitutability of the components of bigrams, we use two resources: (i) ELH: *Sinonimoen Kutxa*, a Basque dictionary of synonyms, published by the Elhuyar Foundation (for nouns and verbs, 40,146 word-synomyn pairs); (ii) WN: the Basque version of WordNet[2](68,217 word-synomyn pairs). First, we experimented with both resources on their own, but the results show that in many cases there either was no substitute candidate, or the corpus lacked combinations containing a substitute. In order to ensure a broader coverage, we combined both resources (ELHWN), and we expanded the set of substitutes including the siblings retrieved from Basque WordNet (ELHWNexpand).

To calculate flexibility, we experiment with the two measures described in section 2.4: z-score and KL-div based R.

3.4 Combining Knowledge Sources Using Machine Learning

We use some ML methods included in the *Weka* toolkit (Hall et al., 2009) in order to combine results obtained in experiments using single knowledge sources (described in section 3.3). The values

[1] http://infomap-nlp.sourceforge.net/

[2] http://ixa2.si.ehu.es/cgi-bin/mcr/public/wei.consult.perl

of the different measures obtained in those experiments were set as features.

We have selected five methods corresponding to different kind of techniques which have been used successfully in this field: Naive Bayes, C4.5 decision tree (j48), Random Forest, SVM (SMO algorithm) and Logistic Regression. Test were carried out using either all features, the features from each type of knowledge, and some subsets, obtained after manual and automatic selection. Following Fazly and Stevenson (2007), verbs are also included as features.

Since, as we will see in section 3.5, the amount of instances in the evaluation dataset is not very high (1,145), cross-validation is used in the experiments for model validation (5 folds). In the case of automatic attribute selection, we use *AttributeSelectedClassifier*, which encapsulates the attribute selection process with the classifier itself, so the attribute selection method and the classifier only see the data in the training set of each fold.

3.5 Evaluation

3.5.1 Reference Dataset and Human Judgments

As an evaluation reference, we use a subset of 1,200 combinations selected randomly from a extracted set of 4,334 bigrams, that is the result of merging the 2,000-best candidates of each AM ranking from the w = ±1 and $f > 30$ extraction set.

The subset has been manually classified by three lexicographers into idioms, collocations and free combinations. Annotators were provided with an evaluation manual, containing the guidelines for classification and illustrative examples.

The agreement among evaluators was calculated using Fleiss' κ. We obtained a value of 0.58, which can be considered moderate, close to fair, agreement. Although this level of agreement is relatively low when compared to Krenn et al. (2004), it is comparable to the one reported by Pecina (2010), who attributed his "relatively low" value to the fact that "the notion of collocation is very subjective, domain-specific, and also somewhat vague." Street et al. (2010) obtain quite low inter-annotator agreement for annotation of idioms in the ANC (American National Corpus). Hence, we consider that the level of agreement we have achieved is acceptable.

For the final classification of the evaluation set, cases where agreement was two or higher were automatically adopted, and the remaining cases were classified after discussion. We removed 55 combinations that did not belong to the NV category, or that were part of larger MWEs. The final set included 1,145 items, out of which 80 were idioms 268 collocations, and 797 free combinations.

3.5.2 Procedure

In order to compare the results of the individual techniques, we based our evaluation on the rankings provided by each measure. If we were to have an ideal measure, the set of bigram categories ('id', 'col' and 'free') would be an ordered set, with 'id' values on top of the ranking, 'col' in the middle, and 'free' at the bottom. Thus, the idea is to compute the distance between a rank derived from the ideally ordered set, which contains a high number of ties, and the rank yielded by each measure. To this end, we use Kendall's τ_B as a rank-correlation measure. Statistical significance of the Kendall's τ_B correlation coefficient is tested with the Z-test. The realistic topline, yielded by a measure that ranks candidates ideally, but without ties, would be 0.68.

In addition, average precision values (AP) were calculated for each ranking.

In the case of association measures, similarity measures applied to VSM, and measures of flexibility, the bigrams were ranked by means of the values of the corresponding measure. In the case of experiments with Lemur, the information used to rank the bigrams consisted of the positions of the documents corresponding to each member of the bigram in the document list retrieved ('rank' in Table 1). For the experiments in which the context sentences have been distributed in different documents, average positions were calculated and weighted, in relation to the amount of documents for each bigram analysis ('rank_weight'). The total number of documents in the list (or 'hits') is weighted in the same manner ('hit_rel').

When using ML techniques, several measures provided by Weka were analyzed: percentage of Correctly Classified Instances (CCI), F-measures for each class (id, col, free), Weighted Average F-measure and Average F-measure.

	measure	τ_B	AP_MWE	AP_id	AP_col
	random rank	(-0.02542)	0.30879	0.0787	0.23358
AM	f	0.18853	0.43573	0.07391	0.37851
	t-score	0.19673	0.45461	0.08442	0.38312
	log-likelihood	0.15604	0.42666	0.10019	0.33480
	PMI	(-0.12090)	0.25732	0.08648	0.18234
	chi-squared	(-0.03699)	0.30227	0.11853	0.20645
DS	R$_{BR}$_NV (MI_-50%)	0.27034	0.47343	0.21738	0.30519
	R$_{W1}$(2000_MI_f3_50%)	0.26206	0.47152	0.19664	0.30967
	L1_Indri_rankNV	0.31438	0.53536	0.22785	0.35299
	L1_KL_rankNV	0.29559	0.51694	0.23558	0.33607
	L2_Indri_hit_rel_NV	**0.32156**	**0.56612**	0.29416	0.35389
	L2_KL_hit_rel_NV	0.30848	0.55146	**0.31977**	0.33241
	L2_Indri_rankN_weight	0.21387	0.45567	0.26148	0.28025
	L2_Indri_rankV_weight	0.31398	0.55208	0.12837	**0.43143**
MSFlex	H_{rel}_Det	0.07295	0.38995	0.12749	0.27704
	H_{rel}_PostAdj	(-0.05617)	0.31673	0.04401	0.29597
	H_{rel}_PreAdj	0.11459	0.38561	0.09897	0.29223
	H_{rel}_Rel	0.09115	0.40502	0.12913	0.29012
	H_{rel}_Num	0.11861	0.43381	0.13387	0.31318
	H_{rel}_ord	(0.02319)	0.31661	0.08124	0.24052
	CPMI (components)	0.05785	0.41917	0.12630	0.30831
LFlex	R_{nv}_ELHWN	(0.08998)	0.36717	0.07521	0.29896
	R_{vn}_ELHWN	(0.03306)	0.31752	0.08689	0.24369
	z-score_V_ELHWNexpand	0.10079	0.35687	0.12232	0.25019
	z-score_N_ELHWNexpand	0.08412	0.35534	0.07245	0.29005

Table 1: Kendall's τ_B rank-correlations relative to an ideal idiomaticity ranking, obtained by different idiomaticity measures. Non-significant values of τ_B in parentheses ($p > 0.05$). Average precisions for MWEs in general, and specific values for idioms and collocations.

4 Experimental Results

4.1 Single Knowledge Experiments

The results for Kendall's τ_B and AP for MWEs and separate AP values for idioms and collocations are summarized in Table 1 (only the experiments with the most noteworthy results are included).

The best results are obtained in the Lemur experiments, most notably in the Lemur_2 type, using either Indri or KL-div indexes. In the MWE rankings, measures of the R-value type only slightly outperform AMs.

In the case of idioms, DS measures obtain significantly better ranks than the other measures. Idioms being the least compositional expressions, his result is expected, and supports the hypothesis that semantic compositionality can better be characterized using measures of DS than using AMs.

Regarding collocations, no such claim can be made, as the AP values for t-score and f outperform DS values, with a remarkable exception: the best AP is obtained by an Indri index that compares the semantic similarity between the verb in combination with the noun and the verb in contexts without the noun (L2_Indri_rankV_weight), accordingly with the claim that the semantics of the verb contribute to the semicompositionality of collocations. By contrast, the corresponding measure for the noun (L2_Indri_rankN_weight) works quite a bit better with idioms than the previous verb measure.

Figure 1 shows the precision curves for the extraction of MWEs by the best measure of each component of idiomaticity.

In Figure 2 and 3, we present separately the preci-

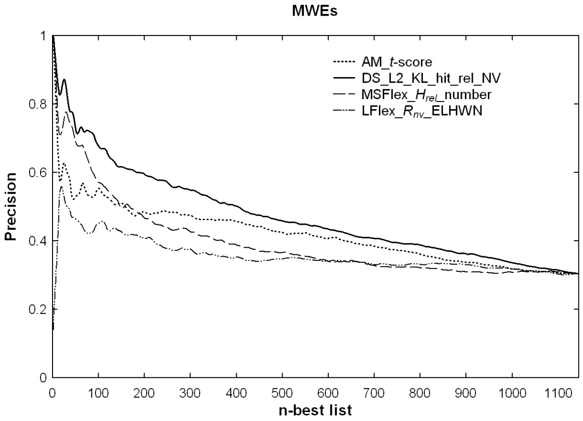

Figure 1: Precision results for the compositionality rankings of MWEs.

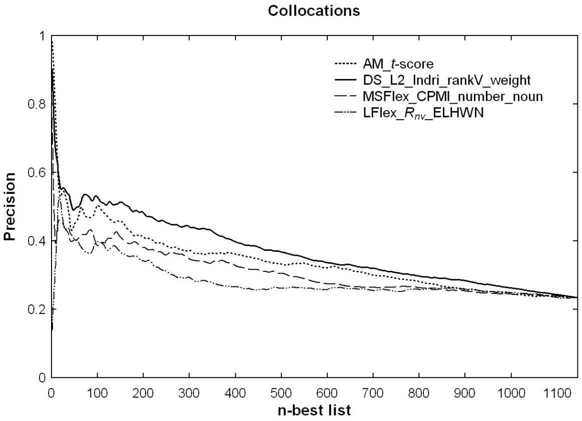

Figure 3: Precision results for the compositionality rankings of collocations.

sion curves for idioms and collocations. We plot the measures with the best precision values.

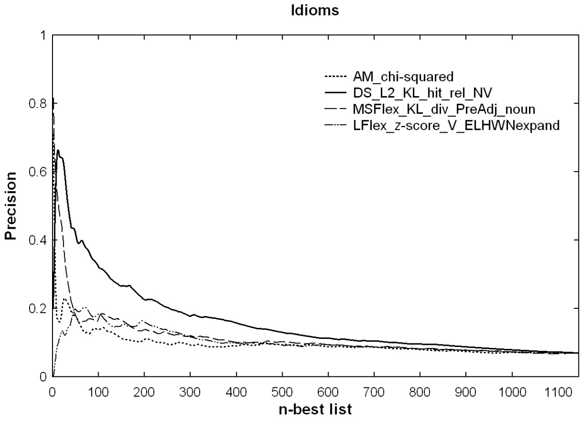

Figure 2: Precision results for the compositionality rankings of idioms.

Regarding the precision for collocations in Figure 3, the differences are not obviously significant. Even though the DS measure has the better performance, precision values for the t-score are not too much lower, and the t-score has a similar performance at the beginning of the ranking (n $<$ 150).

4.2 Machine Learning Experiments

We report only the results of the three methods with the best overall performance: Logistic Regression (LR), SMO and RandomForest (RF).

In Table 2, we present the results obtained with datasets containing only DS attributes (the source of knowledge with the best results in single experiments); datasets containing all features corresponding to the four properties of idiomaticity; and datasets obtained adding the verb of the bigram as a string-type attribute.

As the figures show, it is difficult to improve the results obtained using only DS. The results of SMO are better when the features of the four components of idiomaticity are used, and even better when the verb is added, especially for idioms. The verb causes the performance of RF be slightly worse; in the case of LR, it generates considerable noise.

It can be observed that the figures for LR are more unstable. Using SMO and RF, convergence does not depend on how many noisy variables are present (Biau, 2012). Thus, feature selection could improve the results when LR is used.

In a complementary experiment, we observed the impact of removing the attributes of each source of knowledge (without including verbs). The most evident result was that the exclusion of LFlex features contributes the most to improving F. This was an expected effect, considering the poor results for LFlex measures described in section 4.1. More interesting is the fact that removing MSFlex features had a higher negative impact on F than not taking AMs as features.

Table 3 shows the results for two datasets generated through two manual selection of attributes: (1) manual_1: the 20 attributes with best AP average results; and (2) manual_2: a manual selection of the attributes from each knowledge source with the best AP_MWE, best AP_id and best AP_col. The third

Features	Method	CCI	F_id	F_col	F_free	F_W.Av.	F_Av.
DS	LR	72.489	0.261	0.453	0.838	0.707	0.517
	SMO	74.061	0.130	0.387	0.824	0.575	0.447
	RF	71.441	0.295	0.440	0.821	0.695	0.519
all idiom. properties	LR	71.703	0.339	0.514	0.821	0.716	0.558
	SMO	**76.507**	0.367	0.505	**0.857**	0.740	0.576
	RF	74.498	0.323	0.486	0.844	0.724	0.551
all + verb	LR	60.000	0.240	0.449	0.726	0.627	0.472
	SMO	75.808	**0.400**	**0.540**	0.848	**0.744**	**0.596**
	RF	74.061	0.243	0.459	0.846	0.713	0.516

Table 2: Results of Machine Learning experiments combining knowledge sources in three ways: (i) DS: distributional similarity features; (ii) knowledge related to the four components of idiomaticity (AM+DS+MSFlex+LFlex); (iii) previous features+verb components of bigrams.

section presents the results obtained with *AttributeSelectedClassifier* using *CfsSubsetEval* (CS) as evaluator[3] and *BestFirst* (BS) as search method. Looking at the results of the selection process in each fold, we saw that the attributes selected in more than 2 folds are 36: 1 AM, 20 from DS, 7 from MSFlex, 1 from LFlex and 7 verbs.

Features	Method	F_W.Av.	F_Av.
manual_1	LR	0.709	0.525
	SMO	0.585	0.304
	RF	0.680	0.485
manual_2	LR	0.696	0.518
	SMO	0.581	0.286
	RF	0.688	0.519
CS-BF	LR	**0.727**	**0.559**
	SMO	0.693	0.485
	RF	0.704	0.531

Table 3: F Weighted average and F average results for experiments using: (1) the 20 attributes with best AP average results; (2) a manual selection of the 3 best attributes from each knowledge source; and (3) *AttributeSelectedClassifier* with automatic attribute selection using *CfsSubsetEval* as evaluator and *BestFirst* as search method

The results show that, for each method, automatic selection outperforms the two manual selections. Most of the attributes automatically selected are DS measures, but it is interesting to observe that MSFlex and the verb slot contribute to improving the results. Using automatic attribute selection and LR, the results are close to the best figure of F_W.Av. using SMO and all the features (0.727 vs 0.744).

5 Discussion

The most important conclusions from our experiments are the following:

- In the task of ranking the candidates, the best results are obtained using DS measures, and, in particular, Indri and KL-div in L2 experiments. This is true for both type of MWEs, and is ratified in ML experiments when automatic attribute filtering is carried out. It is, however, particularly notable with regard to idioms; in the case of collocations, the difference between the performance of DS and that of and MS and AM were not that significant.

- MSFlex contributes to the classification task when used in combination with DS, but get poor results by themselves. The most relevant parameter MSFlex is number inflection.

- SMO is the most precise method when a high amount of features is used. It gets the best overall F-score. The other methods need feature selection to obtain similar results.

- Automatic attribute selection using CS-BF filter yields better results than manual selections. The method that takes the most advantage is LR, whose scores are little bit worse than those of SMO using the whole set of attributes.

[3] http://wiki.pentaho.com/display/DATAMINING/CfsSubsetEval

Some of these conclusions differ from those reached by earlier works. In particular, the claims in Fazly and Stevenson (2007) and Van de Cruys and Moirón (2007) that syntactic as well as lexical flexibility outperform other techniques of MWE characterization are not confirmed in this work for Basque. Some hypothesis could be formulated to explain those differences: (1) Basque idioms could be syntactically more flexible, whereas some free combinations could present a non-negligible level of fixedness; (2) Basque, especially in the journalistic register, could be sociolinguistically less fixed than, say, English or Spanish; thus, the lexical choice of the collocate could be not so clearly established; (3) the Basque lexical resources to test substitutability could have insufficient coverage; and (4) Fazly and Stevenson (2007) use the cosine for DS, a measure which in our experiments is clearly below other measures. Those hypotheses require experimental testing and deeper linguistic analysis.

6 Conclusions and Future Work

We have presented an in-depth analysis of the performance of different features of idiomaticity in the characterization of NV expressions, and the results obtained combining them using ML methods. The results confirm the major role of DS, especially, as expected, in the case of idioms. It is remarkable that the best results have been obtained using Lemur, an IR tool. ML experiments show that other features contribute to improve the results, especially some aspects of MSFlex, the verb of the bigram and, to a more limited extent, AMs. The performance of DS being the best one for idioms confirm previous research on other languages, but MSFlex and LFlex behave below the expected. The explanations proposed for this issue require further verification.

We are planning experiments using these techniques for discriminating between literal and idiomatic occurrences of MWEs in context. Work on parallel corpora is planned for the future.

Acknowledgments

This research was supported in part by the Spanish Ministry of Education and Science (TACARDI-TIN2012-38523-C02-011) and by the Basque Government (Berbatek project, Etortek-IE09-262; KONBITZ project, Saiotek 2012). Ainara Estarrona and Larraitz Uria (IXA group) and Ainara Ondarra and Nerea Areta (Elhuyar) are acknowledged for their work as linguists in the manual evaluation. Maddalen Lopez de la Calle and Iñaki San Vicente (Elhuyar) and Oier Lopez de la Calle (IXA group) have contributed with their expertise to the design of the experiments with Lemur and Infomap. Finally, special thanks goes to Olatz Arregi (IXA group) for having guided us in the experiments with Weka, and to Yosu Yurramendi from the University of the Basque Country, for his advice on the statistics in the evaluation step.

References

Allan, J., J. Callan, K. Collins-Thompson, B. Croft, F. Feng, D. Fisher, J. Lafferty, L. Larkey, T. Truong, P. Ogilvie, et al. (2003). The Lemur Toolkit for language modeling and information retrieval.

Baldwin, T., C. Bannard, T. Tanaka, and D. Widdows (2003). An empirical model of multiword expression decomposability. In *Proceedings of the ACL 2003 workshop on Multiword expressions: analysis, acquisition and treatment-Volume 18*, pp. 96.

Baldwin, T. and S. Kim (2010). Multiword expressions. *Handbook of Natural Language Processing, second edition. Morgan and Claypool.*

Banerjee, S. and T. Pedersen (2010). The design, implementation, and use of the Ngram Statistics Package. *Computational Linguistics and Intelligent Text Processing*, 370–381.

Bannard, C. (2007). A measure of syntactic flexibility for automatically identifying multiword expressions in corpora. In *Proceedings of the Workshop on a Broader Perspective on Multiword Expressions*, pp. 1–8.

Berry-Rogghe, G. (1974). Automatic identification of phrasal verbs. *Computers in the Humanities*, 16–26.

Biau, G. (2012). Analysis of a random forests model. *The Journal of Machine Learning Research 98888*, 1063–1095.

Biemann, C. and E. Giesbrecht (2011). Distributional semantics and compositionality 2011:

Shared task description and results. *Workshop on Distributional semantics and compositionality 2011. ACL HLT 2011*, 21.

Church, K. and P. Hanks (1990). Word association norms, mutual information, and lexicography. *Computational linguistics 16*(1), 22–29.

Evert, S. (2005). *The statistics of word cooccurrences: Word pairs and collocations*. Ph. D. thesis, University of Stuttgart.

Fazly, A. and S. Stevenson (2007). Distinguishing subtypes of multiword expressions using linguistically-motivated statistical measures. In *Proceedings of the Workshop on A Broader Perspective on Multiword Expressions*, pp. 9–16. Association for Computational Linguistics.

Granger, S. and M. Paquot (2008). Disentangling the phraseological web. *Phraseology. An interdisciplinary perspective*, 27–50.

Gurrutxaga, A. and I. Alegria (2011). Automatic extraction of NV expressions in Basque: basic issues on cooccurrence techniques. *Proc. of the Workshop on Multiword Expressions. ACL HLT 2011*, 2–7.

Gurrutxaga, A. and I. Alegria (2012). Measuring the compositionality of nv expressions in basque by means of distributional similarity techniques. *LREC2012*.

Hall, M., E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten (2009). The weka data mining software: an update. Volume 11, pp. 10–18. ACM.

Katz, G. and E. Giesbrecht (2006). Automatic identification of non-compositional multi-word expressions using latent semantic analysis. In *Proceedings of the Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*, pp. 12–19. Association for Computational Linguistics.

Krenn, B., S. Evert, and H. Zinsmeister (2004). Determining intercoder agreement for a collocation identification task. In *Proceedings of KONVENS*, pp. 89–96.

Lin, D. (1999). Automatic identification of non-compositional phrases. In *Proceedings of the 37th annual meeting of the ACL*, pp. 317–324. Association for Computational Linguistics.

Lin, J., S. Li, and Y. Cai (2008). A new collocation extraction method combining multiple association measures. In *Machine Learning and Cybernetics, 2008 International Conference on*, Volume 1, pp. 12–17. IEEE.

Oronoz, M., A. D. de Ilarraza, and K. Gojenola (2010). Design and evaluation of an agreement error detection system: testing the effect of ambiguity, parser and corpus type. In *Advances in Natural Language Processing*, pp. 281–292. Springer.

Pecina, P. (2010). Lexical association measures and collocation extraction. *Language resources and evaluation 44*(1), 137–158.

Schone, P. and D. Jurafsky (2001). Is knowledge-free induction of multiword unit dictionary headwords a solved problem. In *Proc. of the 6th EMNLP*, pp. 100–108. Citeseer.

Seretan, V. (2011). *Syntax-Based Collocation Extraction*. Text, Speech and Language Technology. Dordrecht: Springer.

Street, L., N. Michalov, R. Silverstein, M. Reynolds, L. Ruela, F. Flowers, A. Talucci, P. Pereira, G. Morgon, S. Siegel, et al. (2010). Like finding a needle in a haystack: Annotating the american national corpus for idiomatic expressions. In *Proc. of LREC'2010*.

Van de Cruys, T. and B. Moirón (2007). Semantics-based multiword expression extraction. In *Proceedings of the Workshop on A Broader Perspective on Multiword Expressions*, pp. 25–32. Association for Computational Linguistics.

Venkatapathy, S. and A. Joshi (2005). Measuring the relative compositionality of verb-noun (vn) collocations by integrating features. In *Proceedings of HLT/EMNLP*, pp. 899–906. Association for Computational Linguistics.

Wulff, S. (2010). *Rethinking Idiomaticity*. Corpus and Discourse. New York: Continuum International Publishing Group Ltd.