Applying Productive Derivational Morphology to Term Indexing of Spanish Texts

Jesús Vilares[1], David Cabrero[2], and Miguel A. Alonso[1]

[1] Departamento de Computación, Universidad de La Coruña
Campus de Elviña s/n, 15071 La Coruña, Spain
jvilares@mail2.udc.es, alonso@dc.fi.udc.es
http://coleweb.dc.fi.udc.es/
[2] Escuela Superior de Ingeniería Informática, Universidad de Vigo
Edificio Politécnico, As Lagoas, 32004 Orense, Spain
cabrero@uvigo.es

Abstract. This paper deals with the application of natural language processing techniques to the field of information retrieval. To be precise, we propose the application of morphological families for single term conflation in order to reduce the linguistic variety of indexed documents written in Spanish. A system for automatic generation of morphological families by means of Productive Derivational Morphology is discussed. The main characteristics of this system are the use of a minimum of linguistic resources, a low computational cost, and the independence with respect to the indexing engine.

1 Introduction

Spanish is a very rich language in its lexis and its morphology. This implies that Spanish has a great productivity and flexibility in its word formation mechanisms by using productive morphology, preferring derivation to other mechanisms. So it could be interesting to use morphological families for single word terms conflation.

We can define a *morphological family* as a set of words obtained from the same morphological root through derivation mechanisms. It is expected that a basic semantic relationship will remain between the words. To obtain regular word formation patterns, the contribution of generative phonology and transformational-generative grammar let us to speak seriously of 'rules of formation'. Though this paradigm is not complete, it is a great advance in this area, and allows us to implement an automatic system for generation of morphological families with an acceptable degree of completeness and correction.

At this point, we must face one of the main problems of Natural Language Processing (NLP) in Spanish, the lack of available resources. Large tagged corpora, treebanks and advanced lexicons are not available. We will try to overcome these limitations by means of the development of a system using as few resources as possible and confronting this task from the lexical level.

Finally, we will use this system for *single word term conflation* in Information Retrieval (IR) tasks. We will look for simplicity, employing a minimum of linguistic resources (only a tagger which will provide both the part of speech and the lemma), and reducing as much as possible the computational cost. Furthermore, the system will be totally independent from the indexing engine.

2 Morphology and Word Formation

A morpheme may be defined as a 'minimal distinctive unit of grammar', a subunit of the 'word', which in grammatical terms cannot be meaningfully further subdivided [7]. Inflectional morphemes represent grammatical concepts such as gender —*tonto* (silly man), *tonta* (silly woman)—, person, mood, or time and tense —*canté* (I sang), *cantemos* (let's sing)—. On the other hand, we will speak of derivational morphemes when they effect a semantic change on the base and, often, also effecting a change of syntactic class —*aburrir* (to bore), *aburrimiento* (boredom)—. The common remaining element is known as the lexical morpheme or stem.

Morphemes preposed to the base are prefixes, and those postposed are suffixes. Additionally, Spanish affixes are conventionally considered to include infixes, elements which appear internally in the derivational structure —*humareda*, (cloud of smoke)—.

Traditionally, word formation has been divided into compounding and derivation. Compounding involves the combination of independent lexical units —*romper* (to break) + *corazón* (heart) $\rightarrow$ *rompecorazones* (heartbreaker)— and we speak of derivation when one of the components cannot stand on its own as an independent lexeme —*Marx* (Marx) $\xrightarrow{-ismo}$ *marxismo* (Marxism)—, even if it bears significant semantic content —the suffix *-ismo* means "ideology, party, doctrine"—. In either case we are dealing with morphological procedures, the conjoining of individual morphemes or groups of morphemes into larger units forming complex lexemes. These new lexemes we obtain, can be, in their turn, bases for new word formation.

The morphemic structure of the word is fundamental to the analysis of the procedures of word formation. However, we must take into account that in Romance languages, Spanish in particular, it is the word itself rather than any of its morphemic components which produces the derivation. Most Spanish words have a structure such as this: *lexical morpheme + prefix or suffix.*

In addition to compounding and derivation, Spanish is characterized by the frequency of *parasynthesis*, involving the simultaneous prefixation and suffixation of the base lexeme:

$$rojo \text{ (red) } \xrightarrow{en-\ -ecer} enrojecer \text{ (to turn red)}$$

$$alma \text{ (soul) } \xrightarrow{des-\ -ado} desalmado \text{ (heartless)}$$

A critical phenomenon is that many derivational morphemes have variable forms (*allomorphs*), sometimes phonologically determined, at others lexically imposed by convention or etymology:

in-/im-	<u>*in*</u>*necesario* (unnecessary)	<u>*im*</u>*batible* (unbeatable)
des-/dis-	<u>*des*</u>*cosido* (unstitched)	<u>*dis*</u>*culpar* (to forgive)

3 Derivational Mechanisms

In Spanish the basic derivational mechanisms are: *parasynthesis*, *prefixation*, *emotive suffixation*, *non-emotive suffixation* and *back formation*. Parasynthesis and prefixation have been introduced in the previous section. Emotive suffixation semantically alters the base in some sort of subjective emotional way (smallness, endearment, impressiveness, etc.). Non-emotive suffixation and back formation, which are the basis of the system, are discussed in detail below.

3.1 Non-emotive Suffixation

They constitute the general body of suffixes which are considered to be objective in their application, to change the meaning of the base fundamentally rather than marginally, often to have the capacity to effect a change of syntactic category.

The non-emotive suffix repertoire of Spanish is made up of several hundred derivational morphemes whose inventory is not fixed. Furthermore, there are constraints, expansions and changes of all kinds.

One of the problems we find because of the huge number of existent suffixes, is their classification. The criterion we have used is double. The first subdivision we have made is according to the grammatical class of the derivative; so, we have three processes: nominalization to obtain names —the most common—, adjectivization to obtain adjectives and verbalization to form verbs. The second criterion we have used is the grammatical class of the base: denominals (from names), deadjectivals (from adjectives), and deverbals (from verbs).

From a semantic point of view all suffixes are meaningful in the sense that the meaning of the derivative is always different from that of the base. However, most suffixes are polysemic:

$$\textit{muchach}\underline{\textit{ada}}\begin{cases}\text{group } (\textit{group of youths})\\ \text{typical action of the base word } (\textit{childish prank})\end{cases}$$

3.2 Back Formation

This phenomenon is extremely important in contemporary Spanish as a morphological procedure of derivation in *deverbal nominalization*. Instead of increasing the number of syllables of the base, as normally happens is suffixation, it causes a truncation, attaching only a vowel —‘a’, ‘e’ or ‘o’— to the verb stem:

tom<u>*ar*</u> (to take) → *tom*<u>*a*</u> (taking)
altern<u>*ar*</u> (socialize) → *altern*<u>*e*</u> (socializing)
dañ<u>*ar*</u> (to damage) → *dañ*<u>*o*</u> (damage)

4 Phonological Conditions

We must not forget the analysis of the phonological conditions which dominate the word formation process, because every morphological operation involves a phonological alteration of the base. These transformations may be regular or seemingly irregular:

a) *responder* (to reply) → *respondón* (insolent)
vencer (to defeat) → *invencible* (invincible)

b) *pan* (bread) → *panadero* (baker)
agua (water) → *acuatizar* (to land on the sea)

In a) suffixes are attached to the bases or their stems and conjoined in a morphologically regular way, with predictable morphological and phonological results, whereas in b) the morphology, superficially at least, seems irregular in that the output of the derivation is not that we expect (**pandero, *agüizar*). Morphology acts in a complex way in such examples. Although some of those phenomenons seem not to be regular, they are common enough to be forcibly included in any rigorous theory of Spanish lexical morphology.

The attempt to explain such situations has led to the postulation of 'readjustment' rules. Aronoff [7] subdivides these rules into two groups: rules of allomorphy, referred to as allomorphical phenomenons, and rules of truncation, where suffixation requires elimination of a previously existing suffix prior to attaching the new suffix:[1]

pan	→ *panadero*	insertion of /aδ/ separating stem and suffix
/pan/	→ /panaδero/	
agua	→ *acuatizar*	conversion of sonorant /γ/ to obstruent /k/ and
/aγua/	→ /akwatiθar/	insertion of obstruent /t/ before infinitive morpheme

The concrete phonological conditions we have considered in this work are detailed in Sect. 6.

5 Rules and Constraints

Even the most productive patterns of Spanish word formation are subject to constraints. The degree of acceptation of a derivative, well-formed from a morphological point of view, is impossible to be predicted only from its form.

However, if we compare the constraints applied in Spanish and those applied in other languages, we can see the great flexibility of Spanish derivational morphology. If we compare it, for example, with English or French (also a Romance language), we can notice that they are highly inflecting types of languages, but their word formation procedures are not so productive as those of Spanish, particularly with regard to suffixation [7], as is shown in Tables 1 and 2.

Table 1. Word formation in English vs. Spanish

English	Spanish	English	Spanish
fist	*puño*	orange	*naranja*
dagger	*puñal*	orange tree	*naranjo*
stab	*puñalada*	orange grove	*naranjal*
punch	*puñetazo*	orangeade	*naranjada*
fistful	*puñado*	orange coloured	*anaranjado*
to grasp	*empuñar*	orange seller	*naranjero*
to stab	*apuñalar*	blow with an orange	*naranjazo*
sword hilt	*empuñadura*	small orange	*naranjita*

Table 2. Word formation in French vs. Spanish

French	Spanish	French	Spanish
coup de pied	*patada* (kick)	*tête*	*cabeza* (head)
petit ami	*amiguito* (little friend)	*coup de tête*	*cabezazo* (butt)
mal de mer	*mareo* (seasickness)	*incliner la tête*	*cabecear* (to nod)
salle à manger	*comedor* (dining room)	*chef d'èmeute*	*cabecilla* (ringleader)
belle-mère	*suegra* (mother-in-law)	*tète de lit*	*cabecera* (head of a bed)
coup à la porte	*aldabonazo* (bang)	*de tète grosse*	*cabezudo* (big-headed)

We can notice with these examples that Spanish prefers derivation rather than other mechanisms, such as happens in French or English. So, it would be very interesting to use morphological families for single word term conflation.

6 A System for Automatic Generation of Morphological Families

We have incorporated the theoretical basis that we have studied up till now into a computer system for automatic generation of morphological families. The linguistic resources required are minimal, only an incoming lexicon of Spanish, each entry formed by a word, a part of speech tag and the corresponding lemma.

As a first step, the system takes the lexicon and obtains and classifies the lemmas of the words which concentrate the semantic information of a text: names, verbs and adjectives. Classification is needed because the derivational procedures vary depending on the category of the base and on the category of the derivative.

Once we have obtained the set of lemmas, the morphological families are created by applying productive derivational morphology and readjustment rules. The current system deals with *non-emotional suffixation*, *back formation* and some marginal cases of *parasynthesis* in *verbalization*.

[1] phonetical transcription of a word is written surrounded by slash symbols.

Description of the Algorithm. A running example corresponding to the morphological family $\{$*rojo* (red), *enrojecer* (to turn red)$\}$ generated by the base term *rojo* will be used to explain the behaviour of the algorithm.

For every non-processed lemma a new morphological family F is created, with this lemma as its only component. In the running example, F=$\{$*rojo*$\}$. At this moment, F is the *active* family. Then, the lemma is pushed onto the stack S which keeps the non-processed terms of the active family. In the case of the running example, S=[*rojo*].

While S is not empty and non-processed elements exist in F, the algorithm performs the following actions:

1. It gets the lemma on the top of the stack and applies any suitable derivation procedures to it depending on its grammatical category. The lexicon is used to check the existence of the derivative. If the derivative is not valid, phonological conditions will be applied until a valid one is obtained. In the running example, *rojo* is popped from the stack (currently $S = []$) and the parasynthesis *en- -ecer* is used to derive *enrojecer*, which is identified as a correct word in the lexicon.
2. If a correct derivative has been obtained:
   (a) If the derivative has not been previously processed, it is attached to F and pushed onto the stack to be processed later. As a result, F =$\{$*rojo, enrojecer*$\}$ and S =[*enrojecer*] in the running example.
   (b) If the derivative has been previously processed and it belongs to a family $F' \neq F$, then F and F' refer to subfamilies of the same morphological family. In such a case all lemmas of the active family F are re-attached to F', and F' becomes the active family. We call this phenomenon *derivative transivity*, and it would happen, for example, if the lemma *enrojecer* were processed before *rojo*. The obtained family would be F' =$\{$*enrojecer*$\}$. Later, the lemma *rojo* would be processed in F, and *enrojecer* would be derived from it. As a consequence, F and F' would be merged, resulting in the family $\{$*rojo, enrojecer*$\}$.

We can notice that this algorithm overgenerates, i.e. it applies all possible suffixes for the category of a given lemma; so, it obtains all morpholically valid derivatives, which are filtered through the lexicon to select the valid ones. We are solving the problem of deciding about the correctness of the derivative only by its form, without considering other aspects.

With regard to back formation, it is supported indirectly by means of derivative transitivity: instead of deriving the name from the verb, we wait until the name is processed and at that time we obtain the verb through denominal verbalisation.

Allomorphy. There are many factors that affect the choice of the proper allomorphic variant for a given suffix. There exist exclusive variants, such as those whose selection depends on the *theme vowel*[2] —e.g. *-amiento* for theme vowel 'a' and *-imiento* for 'e' or 'i'—. There are other non-exclusive suffixes, such as *-ado/-ato* and *-azgo* [7] (popular and archaic variants, respectively), sometimes producing alternative outputs on the same base:

$$\textit{líder}\ \text{(leader)} \rightarrow \begin{cases} \textit{liderato}\ \text{(leadership, popular)} \\ \textit{liderazgo}\ \text{(leathership, cult)} \end{cases}$$

The proper variant to use depends on each particular suffix, and on factors like the theme vowel or the way the base is formed. Therefore, the different possible situations for each particular suffix have been considered separately.

Phonological Conditions. The most important phonological conditions [2, 3, 7] have been considered:

– **Final unstressed vowel deletion:** It is the default behavior of the system. When the system attaches the suffixes to the base, it first deletes the final unstressed vowel of the lemma. If the ending letter of the base is a consonant, it remains. In any case, the original term remains always available. Examples:

$$\begin{array}{l} \textit{arena}\ \text{(sand)} \rightarrow \textit{aren-} \xrightarrow{-oso} \textit{arenoso}\ \text{(sandy)} \\ \textit{temor}\ \text{(fear)} \rightarrow \textit{temor} \xrightarrow{a-\ -izar} \textit{atemorizar}\ \text{(to frighten)} \end{array}$$

– **Cacophony elimination:** Sometimes, when the suffix is attached, two equal vowels go together. To eliminate the resultant cacophony, we fuse them by first detecting this situation. For example:

$$\textit{galanteo}\ \text{(flirting)} \rightarrow \textit{galante-} \xrightarrow{-ería} \textit{galant}\underline{\textit{ee}}\textit{ría} \rightarrow \textit{galantería}\ \text{(gallantry)}$$

– **Theme vowel:** If the term we are processing is a verb, we check whether it ends in *-<u>a</u>r, -<u>e</u>r* or *-<u>i</u>r* to know the theme vowel and thus take it into account when we want to choose the proper variant to use. Such an example is *-miento, -amiento, -imiento* or *-mento*, where *-amiento* is only used when the theme vowel is 'a', while *-imiento* is used with 'e' or 'i':

$$\begin{array}{c} \textit{alz}\underline{\textit{ar}}\ \text{(to lift)} \rightarrow \textit{alz-} \xrightarrow{-amiento} \textit{alzamiento}\ \text{(lifting)} \\ \textit{aburr}\underline{\textit{ir}}\ \text{(to bore)} \rightarrow \textit{aburr-} \xrightarrow{-imiento} \textit{aburrimiento}\ \text{(boredom)} \end{array}$$

– **Monophthongization of diphthongs:** It is enough to replace the diphthong by the proper form. We manage two different situations:

$$\begin{array}{ll} \text{ie} \rightarrow \text{e} & \textit{diente}\ \text{(tooth)} \xrightarrow{-al} \textit{d}\underline{\textit{ie}}\textit{nt-al} \rightarrow \textit{d}\underline{\textit{e}}\textit{ntal}\ \text{(dental)} \\ \text{ue} \rightarrow \text{o} & \textit{fuerza}\ \text{(strength)} \xrightarrow{-udo} \textit{f}\underline{\textit{ue}}\textit{rz-udo} \rightarrow \textit{f}\underline{\textit{o}}\textit{rzudo}\ \text{(strong)} \end{array}$$

[2] the theme vowel [7] is the conjugationally determined vowel segment which appears in the derivative between stem and suffix —'a' for the first conjugation, 'e' for the second conjugation and 'i' for the third one—.

– **Changes in stress position:** Suffixes generally cause a stress alteration, therefore we must consider that event because in Spanish it may imply spelling changes due to the adding or deletion of accents, which depend on the stress of the word. Most suffixes are stressed, so it is easy to know if we have to add or to delete an accent by simply applying orthographical rules. For example:

$$\textit{europeo}\ \text{(European)}\rightarrow \textit{europe-} \xrightarrow{-ista} \textit{europeísta}\ \text{(pro-European)}$$
$$\textit{novela}\ \text{(novel)}\rightarrow \textit{novel-} \xrightarrow{-ista} \textit{novelista}\ \text{(novelist)}$$

– **Retention of final consonant phonemes:** We can know the original phoneme from the spelling of the original term —e.g. the second ‘c’ in *co<u>c</u>er* (to boil) corresponds to /θ/ and not to /k/—. In the same way, knowing the phoneme, we can know the resultant spelling. For example, the ‘z’ in *cerve<u>z</u>a* corresponds to /θ/ and so

$$\textit{cerve}\underline{\textit{z}}\textit{a}\ \text{(beer)} \rightarrow \textit{cervez-} \xrightarrow{-ería} \textit{cerve}\underline{\textit{c}}\textit{ería}\ \text{(bar)}$$

The phonemes and spelling changes we have considered are the following:

/k/	c → qu
/γ/	g → gü
/γ/	g → gu
/θ/	z → c
/θ/	c → z

– **Ad-hoc rules:** We are referring to varied adjustments such as modifications in the last consonant of the stem, in cases of the kind

$$\textit{conce}\underline{\textit{d}}\textit{er}\ \text{(to concede)} \rightarrow \textit{conce}\underline{\textit{s}}\textit{ión}\ \text{(concession)}$$

They are solved by ad-hoc rules, that is, they manage each particular suffix separately. They are often related to the presence of the dental phonemes /δ/ or /t/.

7 System Evaluation

In order to evaluate the system we have used a lexicon of 995,859 words, 92,125 of them were identified as content word lemmas, finally obtaining 54,243 morphological families. Table 3 shows the number of families of a given size and the number of words stored in families of that given size.

We have taken a random sample of 50 families of 2 or more members, which were manually inspected with the aid of dictionaries to check whether all words really belonged to that family and also to check whether they kept a strong enough semantic relation. We have employed ordinary dictionaries for such a

Table 3. Distribution of the morphological families

	Families		Words	
Size	Number	%	Number	%
1	43,007	79.29	43,007	46.68
2	4,470	8.24	8,940	9.70
3	2,314	4.27	6,942	7.54
4	1,405	2.59	5,620	6.10
5	904	1.67	4,520	4.91
6	501	0.92	3,006	3.26
7	368	0.68	2,576	2.80
8	270	0.50	2,160	2.34
9	223	0.41	2,007	2.18
10+	781	1.43	13,347	14.49
Total	54,243	100	92,125	100

Table 4. Evaluation of the morphological families

correct	79 %
incorrect (2 fam.)	7 %
incorrect (3 fam.)	12 %
incorrect (4+ fam.)	2 %

task because the lexicon we used contained a lot of uncommon words and americanisms. So, those words not included in the dictionaries were considered unusual and were not taken into account for the evaluation. The percentage of these words was about 20% approximately.

A checked family was considered incorrect when one or more of its members were found to really belong to other morphological families. The results we obtained for the sampled families are shown in Table 4, indicating the number of correct families and the number of families containing two or more real families.

We have identified that most mistakes were due to:

– Different families attached through derivational transivity rule. Risk situations are:
  1. Lemmas with similar spelling, specially the shortest ones, for example:
     $$\textit{ano} \text{ (anus) } \xrightarrow{-al} \textit{anal} \xleftarrow{-al} \textit{ana} \text{ (length measure)}$$
  2. Monophthongisation of diphtongs, for example:
     $$\textit{fuel} \text{ (fuel) } \xrightarrow{-\acute{\imath}a} \textit{fol\'{\i}a} \text{ (dance)}$$
  3. Parasynthesis, for example:
     $$\textit{plasta} \text{ (soft mass) } \xrightarrow{a-\ \ -ar} \textit{aplastar} \text{ (to flatten)}$$
– Existence of more than one sense for the same lemma. For example:

rancho (communal meal)
rancho (ranch) $\xrightarrow{-ero}$ *ranchero* (rancher)
rancho (shanty) $\xrightarrow{-ería}$ *ranchería* (shanty town)

– Sense specialization:

golpe (hit) $\xrightarrow{-ador}$ *golpeador* (that hits)
golpe [de estado] (coup d'état) $\xrightarrow{-al}$ *golpista* (participant in a coup d'état)

– Figurative senses, for example:

lince (lynx) $\xrightarrow{-ear}$ *lincear* (to unearth)

There are some procedures to limit the possible mistakes, such as using etymological or semantic information to check whether there is any relation between the derivative and the base lemma. But this would imply increasing the computational costs, and problems would not completely disappear. In the case of employing etymology, there are examples of unrelated words such as *Morfeo* (Morpheus) and *morfina* (morphine) with the same etymologycal origins. In the case of using semantic information, if a word has more than one sense we would need to disambiguate it from the context.

8 Term Indexing Using Morphological Families

Information Retrieval systems conflate the documents before indexing to decrease their linguistic variety by grouping together textual occurrences referring to similar or identical concepts exploiting graphical similarities, thesaurus, etc. [1, 5, 6]. We will study another way, the use of morphological families.

For this purpose, we will first get the part of speech and the lemmas of the text to be indexed by using a tagger. Next, we will replace each of the obtained lemmas by a representative of its morphological family. We are replacing all lemmas belonging to the same family by the same lemma, its *representative*; therefore, we are representing all its members by a single term.

Three kind of indexing methods have been tested:

1. Indexing of the original document (*pln*).
2. Indexing of the conflated text via lemmatization (*lem*).
3. Indexing of the conflated text via morphological families (*fam*).

The evaluation of an IR system involves the computation of the standard measures of *precision* P and *recall* R, where:

$$P = \frac{\text{number of relevant documents retrieved}}{\text{total number of documents retrieved}}$$

$$R = \frac{\text{number of relevant documents retrieved}}{\text{total number of relevant documents}}$$

The reference document corpus used for testing was composed of 1,378 documents with an average length of 292 words. The topics of the corpus were of a journalistic nature (local, national, international, sports, culture).

We have used SWISH-E [10], a free distribution indexing engine. This software employs a boolean model [1, 6]. This fact is very important because it does not allow partial matches, so recall will be considerably decreased. Though, one of the main advantages of our system is that it is independent from the indexing engine used, because documents are preprocessed before being treated by it; so, there is no problem in changing the indexing engine in the future.

The results we have obtained for global recall and precision for 12 different queries are shown in Table 5 and seem to prove that the application of conflation techniques *[fam,lem]* to reduce the linguistic variety of the documents has lead to a remarkable increase in recall. We can see that the greatest recall is reached using conflation via morphological families *[fam]*. However, this increase in recall also implies a slight decrease in precision. We can also notice that for both measures the worst method is the indexation of the original text *[pln]*. In this last case, precision decreases because no documents are retrieved for some queries due to its sensitivity to inflectional variations (gender, number, time, etc.).

Table 5. Recall and Precision for the different models

	fam	*lem*	*pln*
Recall	0.950	0.738	0.403
Precision	0.693	0.723	0.589

With respect to the evolution of precision vs. recall, Fig. 1 confirms the *pln* model as the worst one. The best behavior corresponds to the *fam* model, except for the segment of recall between 0.15 to 0.33, where the *lem* model is the better one.

9 Conclusion

This paper explores a new approach to retrieval of Spanish texts based on the application of natural language techniques. To be precise, we have investigated the automatic generation of morphological families by applying productive derivational morphology mechanisms, and the exploitation of these families for single term conflation, substituting every term of the document by a representative of its morphological family.

We have looked for simplicity to overcome the limitations associated with the lack of linguistic resources for Spanish, and we have also looked to reduce the computational cost of the system. The solution we have used to solve these problems has been to use as few resources as possible, facing the task from the lexical level. Due to the fact that all families and their representatives have

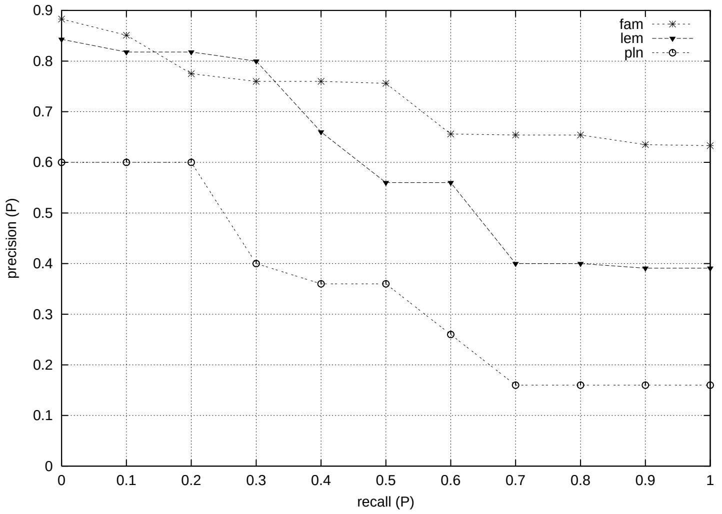

Fig. 1. Precision vs. Recall of the different models

been obtained previously in an independent process, the computational cost is reduced to the cost of the necessary previous lemmatization.

There exist other approaches to the problem of term conflation by using more linguistic resources and increasing the computational cost associated. For example, in [5] a transducer is used for both, morphological analysis and dynamically generating terms of the same derivational family. In [4], the CELEX morphological database[3] has been used to calculate morphological families. In addition, two sources of semantic knowledge for English processing have been applied: the WordNet 1.6 thesaurus [8] and the thesaurus of Microsoft Word97.

The performance of the new approach has been compared to two other indexing methods, indexing of the original text and indexing of the lemmatized text, which obtain worse results in general. However, larger experiments (with a larger collection of texts and a larger set of queries) should be performed to confirm these results.

We are currently extending the system, incorporating more derivational mechanisms (e.g. prefixation), multiword terms [5] and the combination of the retrieved documents by different techniques. We are also extending our tests to different indexing models (e.g. vectorial model[1, 6]) and different indexing engines.

[3] a morphological database for English language in which each lemma is associated with a morphological structure that contains one or more root lemmas.

Acknowledgements

We would like to thank Margarita Alonso, Jean-Cédric Chappelier, Jorge Graña and Manuel Vilares for fruitful discussions. This research has been partially supported by the FEDER of EU (Grant 1FD97-0047-C04-02) and Xunta de Galicia (Grant PGIDT99XI10502B).

References

1. Baeza-Yates, R., Ribeiro-Neto, B.: Modern information retrieval. Harlow, England: Addison-Wesley (1999)
2. Bajo Pérez, E.: La derivación nominal en español. Madrid: Arco Libros, Cuadernos de lengua española (1997)
3. Fernández Ramírez, S.: La derivación nominal. Madrid: Real Academia Española, Anejos del Boletín de la Real Academia Española **40** (1986)
4. Jacquemin, C.: Syntagmatic and paradigmatic representations of term variation Proc. of 37th Annual Meeting of the Association for Computational Linguistics (ACL'99), University of Maryland, USA
5. Jacquemin, C., Tzoukermann, E.: NLP for Term Variant Extraction: A Synergy of Morphology, Lexicon and Syntax. In T. Strzalkowski, editor, Natural Language Processing Information Retrieval, pp. 25–74. Boston: Kluwer (1999)
6. Kowalski, G.: Information retrieval systems : theory and implementation. Boston: Kluwer (1997)
7. Lang, M. F.: Spanish Word Formation: Productive Derivational Morphology in the Modern Lexis. London: Routledge (1990)
8. Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D. and Miller, K. J.: Introduction to WordNet: An On-line Lexical Database. International Journal of Lexicography **3** (1990) 235–244
9. Strzalkowski, T. (editor): Natural language information retrieval. Dordrecht: Kluwer (1999)
10. SWISH-E. http://sunsite.berkeley.edu/SWISH-E/. SWISH-Enhanced - Digital Library SunSITE