Morphological acquisition by Formal Analogy

Jean-François Lavallée and Philippe Langlais
DIRO, University of Montreal
Quebec, Canada
{lavalljf,felipe}@iro.umontreal.ca

Abstract

While most approaches to unsupervised morphology acquisition often rely on metrics based on information theory for identifying morphemes, we describe a novel approach relying on the notion of *formal analogy*, that is, a relation between four forms, such as: `reader` is to `doer` as `reading` is to `doing`. Our assumption is that formal analogies identify pairs of morphologically related words, for instance `reader/reading` and `doer/doing`. Based on this assumption, our approach simply consists in identifying all the formal analogies involving the words in a lexicon. It turned out that for large lexicons, this happens to be a very time consuming task. Therefore, we report our attempts in designing practical systems based on the analogical principle. We applied our systems on the five languages of the shared task. The learning is made in an unsupervised manner, making use of the supplied lexicons only.

General Terms

Formal Analogy, Morphology, Unsupervised Morphology Acquisition

1 Introduction

1.1 Common approach to morphology

In the last few years, two predominant approaches in morphological analysis system seem to have emerged. One of those trends is the usage of a probabilistic approach to find probable segmentations of words. The basic idea is that low predicability of the upcoming letter in a string indicates a morpheme boundary since a morpheme should be seen together with with several different morpheme. This approach has been around for quite some time. Harris [5] described such a system in an article published in 1955. More recently a similar concept has been used by Bernhard [2] and in the system Morfessor [3] who uses perplexity as one of his metrics to validate potential segmentations. The second brand of approaches consists in grouping words into paradigms and removing the common affix of those groups. This approach was prevalent in the Morpho Challenge 2008 competition [11, 15] and will probably still be popular in the 2009 evaluation because of the excellent ranking obtained by Monson's system: `Paramor`.

1.2 Previous application of Formal Analogy

Recently, analogy-based systems gained in popularity as some researchers showed that *analogical learning* based on *formal analogy* can be applied to many canonical problems of computational linguistic. It is especially true for machine translation since the work of Lepage and Denoual [10]. The authors demonstrated that it is possible to translate sentences of a limited domain using only the concept of formal analogy. They reported good results in five translation directions (Japanese, Chinese, Corean, Arabic to English and English to Chinese). Langlais and Patry [8] translated unknown words in French, Spanish and German from/into English using this concept, while Denoual [4] has done the same for English and Japanese. Langlais *et al.* [9] validated that the same principle can be used to translate multi-terms of the medical domain in ten translation directions (French, Finnish, Swedish, Spanish, Russian from and into English).

1.3 Formal Analogy applied to morphology

Some studies on formal analogies are more specifically applied to morphology. Stroppa & Yvon [13] demonstrate that analogical learning can be used in order to recover the lemma as well as an ensemble of characteristics (gender, plural, etc.) of a word. They reported state-of-the-art results for three languages (English, Dutch and German). Hathout [6, 7] reported several studies where morphological families of words are automatically extracted thanks to formal analogies and some semantic resources.

If those studies show the link between formal analogies and morphology, there was no attempt to see whether this concept could lead to a competitive system for the tasks considered in Morpho Challenge. This work aims at filling this gap. The remainder of this paper is as follow. We first give in section 2 a brief recall of the definition of formal analogy we are using in this study and then present the principle of a system based on this concept. As will be described shortly, one issue with our approach is its scalability, therefore, we describe in section 3 some practical systems we devised. We present in section 4 the experimental protocol we followed while developing our systems and the results we obtained. We conclude this work and discuss some future avenues in section 5.

2 The core analogical system

The core idea of the systems we have designed relies on the notion of formal analogy. In this section, we first review this notion, then we present the analogical device, and discuss some practical issues involved.

2.1 Formal analogy

A (formal) *proportional analogy*, or analogy for short, is a relation between four items noted $[x : y = z : t]$ which reads as "$\mathtt{x}$ is to $\mathtt{y}$ as $\mathtt{z}$ is to $\mathtt{t}$". Among proportional analogies, we distinguish *formal analogies*, that is, those we can identify at a graphemic level, such as [*cordially* : *cordial* = *appreciatively* : *appreciative*]. Formal analogies can be defined in terms of factorization. Let $\mathtt{x}$ be a string over an alphabet Σ, a *factorization* of $\mathtt{x}$, noted $f_{\mathtt{x}}$, is a sequence of n *factors* $f_{\mathtt{x}} = (f^1_{\mathtt{x}}, \ldots, f^n_{\mathtt{x}})$, such that $\mathtt{x} = f^1_{\mathtt{x}} \odot f^2_{\mathtt{x}} \odot f^n_{\mathtt{x}}$, where $\odot$ denotes the concatenation operator. After [14] we thus define a formal analogy as:

Definition 1 $\forall (\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}, \boldsymbol{t}) \in {\Sigma^{\star}}^4$, $[x : y = z : t]$ ***iff*** *there exists factorizations* $(f_{\boldsymbol{x}}, f_{\boldsymbol{y}}, f_{\boldsymbol{z}}, f_{\boldsymbol{t}}) \in ({\Sigma^{\star}}^d)^4$ *of* $(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}, \boldsymbol{t})$ *such that,* $\forall i \in [1, d]$, $(f^i_{\boldsymbol{y}}, f^i_{\boldsymbol{z}}) \in \left\{(f^i_{\boldsymbol{x}}, f^i_{\boldsymbol{t}}), (f^i_{\boldsymbol{t}}, f^i_{\boldsymbol{x}})\right\}$. *The smallest* d *for which this definition holds is called the degree of the analogy.*

According to this definition, [*cordially* : *cordial* = *appreciatively* : *appreciative*] is an analogy because we can find a quadruplet of 4-factorizations (factorizations involving 4 factors) as shown in the first column of Figure 1. The second column of this figure also shows that a quadruplet of 2-factorizations also satisfies the definition. This illustrates the *alternations* passively captured by this analogy, that is, `appreciative/cordial` and `ly/`ϵ; the latter one capturing the fact that in English, an adverb can be constructed by appending `ly` to an adjective.

2.2 The analogical device

Although our definition of formal analogy do not enforce a direct correspondence with morphology, a simple inspection of the formal analogies identified within a lexicon shows that morphological

$\mathrm{f}_{cordially}$	$\equiv$	cordia	l	l	y	$\mathrm{f}_{cordially}$	$\equiv$	cordial	ly
$\mathrm{f}_{cordial}$	$\equiv$	cordia	ϵ	l	ϵ	$\mathrm{f}_{cordial}$	$\equiv$	cordial	ϵ
$\mathrm{f}_{appreciatively}$	$\equiv$	appreciative	l	ϵ	y	$\mathrm{f}_{appreciatively}$	$\equiv$	appreciative	ly
$\mathrm{f}_{appreciative}$	$\equiv$	appreciative	ϵ	ϵ	ϵ	$\mathrm{f}_{appreciative}$	$\equiv$	appreciative	ϵ

Figure 1: Two factorizations of the analogy [*cordially* : *cordial* = *appreciatively* : *appreciative*].

information is often captured. Actually, the systems we have designed rely on the assumption that a formal analogy implicitly relates two pairs of forms that are morphologically related. In our running example, `cordially` and `cordial`, as well as `appreciative` and `appreciatively` are morphologically related. This means that in order to participate to task 1 of Morpho Challenge, we need a way to identify formal analogies between words in a lexicon. Formally, given a lexicon of forms $\mathcal{L}$, we seek for:

$$\mathcal{A}(\mathcal{L}) = \{(\mathtt{x}, \mathtt{y}, \mathtt{z}, \mathtt{t}) \in \mathcal{L}^4 \ : \ [x : y = z : t]\}$$

Stroppa [12] describes a dynamic programming algorithm which checks whether a quadruplet of forms $(\mathtt{x}, \mathtt{y}, \mathtt{z}, \mathtt{t})$ is a formal analogy according to the definition we gave. The complexity of this algorithm is in $o(|\mathtt{x}| \times |\mathtt{y}| \times |\mathtt{z}| \times |\mathtt{t}|)$.

2.3 Practical issue

As simple as it seems, it turned out that identifying formal analogies is a very time consuming process. A straightforward implementation would require to check $o(|\mathcal{L}|^4)$ analogies, where $|\mathcal{L}|$ is the number of words in the lexicon. For all but tiny lexicons, this is simply not manageable. In [8], the authors describe some ways to fasten the process. We used the `tree-count` strategy they describe, the details of which are not relevant here. Suffices it to say that this strategy still is too time consuming for the largest lexicons being used here. We roughly estimated that several months of computation would be required for a single desk-computer to acquire all the possible analogies involving words of the Finnish lexicon. Needless to say, we did not have access to such a computation power. Instead, we ran the analogical device over a week period in order to acquire a large set of analogies per language. From 11 (VOWARA) to 52 (TUR) millions of analogies were identified this way. While these figures might seem large at a first glance, it is important to note that this represent only a small part of the analogies that can potentially be identified. Note also that due to the small size of the Arabic lexicons, we managed to collect all the analogies for this language.

3 Systems developed for the shared task

We have developed two families of systems for the Morpho Challenge shared task. The first family encompasses two systems which simply uses the analogies in order to compute a set of `c-rule`s, a notion we will describe shortly. Those `c-rule`s are then used to accomplish the morphological analysis. The second family of systems are pure analogical systems. Since for most of the languages, we did not manage to compute **all** the analogies involving the forms of a given lexicon, the corresponding systems are suffering data-sparsity. Therefore, they were studied as a proof of concept rather than as fully fledged systems. The remainder of this section describes in details both families of systems.

3.1 Cofactor-based systems

COF-GRAPH and COF-FIRST are two systems that are making use of the notion of `c-rule`s in order to build links between morphologically related words. Both are using a structure we call a Word-Relation Tree, or WRT for short. The two systems differ in the way this structure is built.

3.1.1 Cofactor and c-rule

Due to the amount of time needed to compute all analogies, we needed a way to generalize learning from a subset of words to the whole lexicon. This necessity has lead to the introduction of the concept of `c-rule`, a transformation of the notion of *cofactor* introduced in [8] into a rewritting rule. The authors defined the cofactors of a formal analogy $[x : y = z : t]$ as a vector of d alternations $[\langle \mathtt{f}, \mathtt{g}\rangle_i]_{i\in[1,d]}$ where an alternation is defined formally as:

$$\langle \mathtt{f}, \mathtt{g}\rangle_i = \begin{cases} (f_{\mathtt{x}}^{(i)}, f_{\mathtt{z}}^{(i)}) & \text{if } f_{\mathtt{x}}^{(i)} \equiv f_{\mathtt{y}}^{(i)} \\ (f_{\mathtt{y}}^{(i)}, f_{\mathtt{z}}^{(i)}) & \text{otherwise} \end{cases}$$

and d is the degree of the analogy. For instance, the cofactors of the analogy of our running example are: $[(\mathtt{cordial}, \mathtt{appreciative})\ , (\epsilon, \mathtt{ly})]$. Note that the pairs of forms in this definition are not directed, that is, $(\epsilon, \mathtt{ly})$ equals $(\mathtt{ly}, \epsilon)$. A formal analogy captures relations between forms, but the information is only latent and highly lexical. For instance, knowing that [*cordial* : *cordially* = *appreciative* : *appreciatively*] does not tell us anything about [*cordialness* : *appreciativeness* = *cordial* : *appreciative*] or [*passive* : *passively* = *massive* : *massively*]. The notion of cofactor helps in generalizing the information captured by analogies. For instance cofactors such as $(\epsilon, \mathtt{ly})$ or $(\mathtt{ity}, \mathtt{ive})$ abstract away suffixations operations frequently involved in English. At the same time, a cofactor such as $(\mathtt{un}, \epsilon)$ which might capture a prefixation operation in English (*e.g.* `related`/`unrelated`) can wrongly relates forms such as `aunt` and `at` just because one form happens to contain the substring `un`.

Clearly, the generalization offered by a cofactor might introduce some noise if applied blindly. This is the motivation for the concept of `c-rule` we introduced in this work. A `c-rule` is a directed cofactor that is expressed as a rewriting rule $[\alpha \rightarrow \beta]$, where α and β are the two factors of a cofactor, and such as $|\alpha| \geq |\beta|$.[1] As a result, successive applications of a `c-rule` will tend toward a shorter word, our assumption being that shorter words tend to be morphologically simpler. In order to distinguish prefixation and suffixation operations which are very frequent, we add the symbol $\star$ to the left or/and right of the factors to indicate the existence of non ϵ factors. The two `c-rules` $[\star\text{ly} \rightarrow \star\epsilon]$ and $[\text{appreciative}\star \rightarrow \text{cordial}\star]$ are learned from our running example. In the sequel, we note $\mathcal{R}(\mathtt{x})$, the application of the `c-rule` $\mathcal{R}$ on a word `x`. For instance, if $\mathcal{R}$ is $[\star\text{ly} \rightarrow \star\epsilon]$, $\mathcal{R}(\mathtt{elderly})$ equals `elder`.

3.1.2 Extraction of c-rules

From the set of computed analogies, we extract every `c-rule` and their frequency of occurrence. As described in Section 2.3, the amount of analogies generated is enormous and so is the number of `c-rule`. Therefore, we apply a first filter which removes low-frequency `c-rules`.[2]

Relying on counts biases the `c-rules` towards those containing short factors. For instance in English, the `c-rule` $[\text{anti-}\star \rightarrow \epsilon\star]$ is seen 2 472 times, while $[\text{ka}\star \rightarrow \epsilon\star]$, which is likely fortuitous, is seen 13 839 times. To overcome this, we score a `c-rule` $\mathcal{R}$ by its *productivity* $\mathcal{P}$ defined as the ratio $\mathcal{P}(\mathcal{R})=\mathcal{V}/\mathcal{A}$ of the number of time it's application leads to a valid result $\mathcal{V}$ over the number of times it could be applied $\mathcal{A}$. Since in our case we don't have a set of training examples, we estimate productivity by considering every application that leads to a word contained in the lexicon $\mathcal{L}$ a valid result. We consider a rule applicable to a word if the resulting word differs from the input word. Formally, we have $\mathcal{V} \simeq |\{x \in \mathcal{L} : \mathcal{R}(x) \in \mathcal{L}\}|$ and $\mathcal{A} = |\{x \in \mathcal{L} : \mathcal{R}(x) \neq x\}|$. Using productivity, the `c-rule` $[\text{anti-}\star \rightarrow \epsilon\star]$ outweights $[\text{ka}\star \rightarrow \epsilon\star]$ with a score of 0.9490 versus 0.2472.

3.1.3 Word-relation tree construction

Both our systems rely on building a forest of WRT where the roots of those trees are the stems of their descendents. A WRT is a structure where nodes are the words of the lexicon and edges

[1] In case both factors have the same length, alphabetical ordering is used.

[2] `c-rules` occurring less than 20 times are removed.

count	prod
[⋆'s → ⋆ϵ]	[already-⋆ → ϵ⋆]
[⋆s → ⋆ϵ]	[american-⋆ → ϵ⋆]
[⋆s⋆ → ⋆ϵ⋆]	[⋆-backed → ⋆ϵ]
[⋆e⋆ → ⋆ϵ⋆]	[brain-⋆ → ϵ⋆]
[⋆a⋆ → ⋆ϵ⋆]	[⋆-buying → ⋆ϵ]

Table 1: Top-5 `c-rules` ranked by frequency of occurrence (count) and productivity (prod).

correspond to the application of a set of `c-rules` from a child node to its father node. Therefore, the morphological complexity of the words increases further down the tree (see Figure 2). The construction of a WRT is a greedy process, where among all the potential links l between two nodes, we pick the one with the largest score $\mathcal{S}(l)$. The only difference between COF-FIRST and COF-GRAPH comes from how $\mathcal{S}(l)$ is calculated, as described shortly. In both our system, a link is considered if its score $\mathcal{S}(l)$ is higher than a threshold.[3]

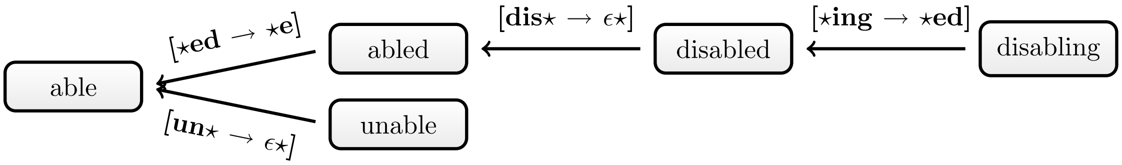

Figure 2: Part of the tree for the stem `able`, note that for COF-GRAPH one arc might be associated to more than one cofactor.

In case of the COF-FIRST, only one `c-rule` $\mathcal{R}$ can be applied from a child node c to its parent f (that is, $\mathcal{R}(c) = f$) and the score of a the link is simply the productivity of $\mathcal{R}$: $\mathcal{S}(l) = \mathcal{P}(\mathcal{R})$. One limitation of this system is that in order for two words to be related by more than one `c-rule`, there must exists a path connecting those words by a single `c-rule`. For instance, in order to relate `able` to `disabled`, a word such as `disable` must exist in the lexicon. To overcome this limitation, we have created a more complex system COF-GRAPH. In this system, the score of a link $l \equiv (c, f)$ is computed by building a directed graph of all the forms that can be obtained by successive applications of any number of `c-rules` to the word in c, as illustrated in Figure 3 for the word `disabled`.

This graph is used in order to compute $\mathcal{S}(l)$. For this, all the paths from c to f are considered. The score of a path is the product of the productivity of each `c-rule` implied. Since several paths might exist between two nodes, we set $\mathcal{S}(l)$ to the sum of the scores of each path. In our example, the score of the link between `disabled` and `able` would be computed by summing the scores of the three paths linking those two nodes.

3.1.4 Extracting morphemes from the Word-Relation tree

Each node in a WRT contains the morphemes of the associated word. In case of the root node, the set of morphemes is a singleton containing the word itself. For inside nodes, the set of associated morphemes is given by the union of the morphemes of its father to the morphemes extracted from all the `c-rule` that created the link. A morpheme is extracted from a `c-rule` by taking the left part of the *equivalent* *`c-rule`* with the highest count. To take one simple example, in Figure 2, the word `disabled` is associated to the set [`dis`, `able`, `ed`] because `abled` is represented by the set [`able`, `ed`] and `dis` is extracted from the associated `c-rule`.

Very often, the `c-rule` labeling an edged induces a segmentation that does not correspond to a morpheme boundary. For instance, in the `c-rule` [⋆bled → ⋆ble], `bled` is not a morpheme.

[3] We set this threshold to 0.35 in this study.

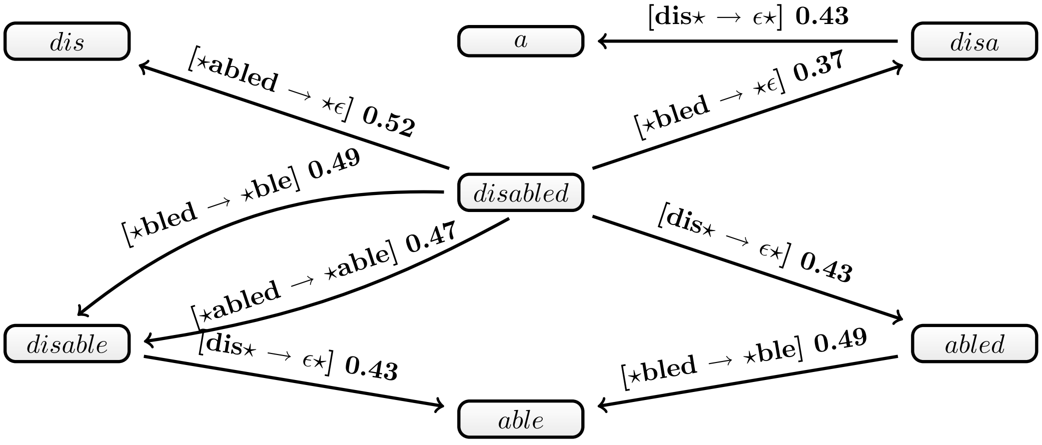

Figure 3: Graph build for the word `disabled` in COF-GRAPH. The most probable link is `disable` with a score of 0.96. With COF-FIRST, `dis` would be selected with a score of 0.52.

This is why we introduced the notion of *equivalent* `c-rule`. A `c-rule` $\mathcal{R}_a$ is equivalent to another `c-rule` $\mathcal{R}_b$ if the rule $\mathcal{R}_a$ always gives the same result than the rule $\mathcal{R}_b$ on the subset of the words on which the rule is applicable as defined in Section 3.1.2. For example, [⋆ed → ⋆e] and [⋆d → ⋆ϵ] are equivalent to [⋆ted → ⋆te]. Note that if the right factor of the equivalent `c-rule` is part of the morphemes of the father, this morpheme won't be included in the morphemes of the child. This is necessary because sometimes intermediate relationships are created in the tree such as the link `disabling/disabled` in Figure 2, which induces the decomposition [`dis`, `able`, `ing`] for the form `disabling` where the form `ed`, which belongs to the decomposition of `disabled` has been removed.

3.2 Pure analogical systems

We implemented two systems making a direct use of the analogies we collected. The first one, named ANA-SEG, uses the factorizations involved in a given analogy in order to segment forms into factors. The second one, named ANA-PAIR, simply links together the related forms in an analogy. It is important to note that because we computed only a small portion of all the analogies, there are many words that these two systems do not treat adequately. In particular, the words for which no analogy is identified are left as is in the final solution, which clearly impacts recall.

3.2.1 ana-seg

Each time a word is involved in an analogy, we can compute its factorization, as explained in section 2. It is therefore possible to maintain a distribution over the *segmentations* computed this way for given word. Figure 4 illustrates the segmentations produced for a few selected words in different languages. For instance, the English word `abolishing` is involved in 21 analogies, leading to 6 different factorizations that are presented in decreasing order of frequency. The ANA-SEG system simply picks the most frequent factorization observed for each word. So in our example `abolish+ing` will be produced for the word `abolishing`.

3.2.2 ana-pair

In this variant, the analogies are checked for related words. This amounts to check whether the first and second words (as in [*reader* : *unreadable* = *doer* : *undoable*]) or the first and the third one (as in [*reader* : *doer* = *unreadable* : *undoable*]) are related. This can be done straightforwardly by comparing the factors of each word. Therefore, for each analogy, we identify two related pairs of words. For a given pair of words (`a`, `b`), we simply add the "morpheme" `a+b` in the entries

abolishing (eng)		abberufen (ger)		abdallardan (tur)	
abolish ing	12	ab berufen	12	abd allardan	17
ab olishing	4	a b b erufen	12	abdallar dan	10
abol ishing	2	abberufe n	10	a b da llardan	9
a bo lishing	1	a b beruf en	6	ab dallardan	6
abolis hing	1	abb erufen	5	abdallar d an	5
abolish in g	1	abberuf en	5	a b da l lardan	4
		ab beruf en	2	ab dallar dan	2
		abbe rufe n	1	abda llardan	2

Figure 4: Factorizations induced by analogy for some words in different languages. Numbers indicate the frequency of a given factorization.

of both `a` and `b`. Figure 5 shows an excerpt of the output produced by ana-pair for the entry `kennzeichneten`. The first "morpheme" indicates that `auszeichneten` and `kennzeichneten` have been related in a given analogy. Clearly, the output produced by ana-pair is far from explaining the constructive morphology of a given word, but allows to evaluate the relevance of the link found by analogy.

auszeichneten+kennzeichneten gekennzeichneten+kennzeichneten kennzeichneten+vorgezeichneten kennzeichneten+ungekennzeichneten bezeichneten+kennzeichneten gekennzeichnete+kennzeichneten kennzeichneten+vorzeichneten bezeichnete+kennzeichn eten kennzeichnete+kennzeichneten aufgezeichneten+kennzeichneten kennzeichnet+kennzeichneten aufzeichneten+kennzeichneten

Figure 5: Output generated by ana-pair for the German word `kennzeichneten`.

4 Experiments

4.1 Corpora and evaluation

Internally, we evaluated variants of our systems on two testbeds. The first one, hereafter mc-ref, is the one provided by the organizers for each language and consists in a small excerpt of between 400 to 700 words by language. Because those references are very small, we also decided to conduct larger-scaled evaluations, making use of Celex [1]. The extraction has been done by generating all possible morphological analysis contained in Celex and removing from the analysis the morphemes that had no equivalent in the Morpho Challenge gold standard. Table 2 describes the resulting references. This latter evaluation named celex-ref in the following, was conducted for English and German, the only two languages in common between Celex and the shared task.

	Lexicon size	Morpheme type	analysis/word	morpheme/analysis	pair/Word
eng	72 628	16 388	1.07	2.15	21.93
ger	311 000	13 102	1.23	3.35	327.75

Table 2: Main characteristics of the celex-ref gold standard. From left to right, the size in word, the number of distinct morpheme, the average amount of possible analysis for a word, the average amount of morpheme in a word, the average number of word that a word can be paired with on the basis that they share at least one morpheme

4.2 Results

The performance of the four systems on the MC-REF task are described in Table 3 while Table 4 shows the result of the same systems on the CELEX-REF task. Globally, we noted that the results on CELEX-REF are coherent with those obtained on MC-REF. On both reference sets, the cofactor-based systems outperformed the pure-analogical ones for all but the arabic language. This observation does not come at a surprise, since the latter systems suffer from the fact that only a small subset of the analogies have been identified, as we already discussed. This clearly impacts recall.

We observe that the performance of our systems is coherent for all languages but the Arabic language where we have a lower F-score due to a really low recall. This might be caused by the size of the arabic lexicon that is more than 10 times smaller than the second smallest lexicon provided, with less than 20 000 words. Since analogical learning somehow relies on the pattern frequency to identify morphemes, several valid morphemes might be overlooked due to their low frequency in the training set. The high precision supports this hypothesis as it shows that what the systems manage to learn from the lexicon is valid but that only a few morphological phenomenon could be identified.

One point to note is that the usage we have made of the extracted analogies is not adapted to some particularities of the Arabic language. This is illustrated by the entry for the word *atawAo* of the vowelized gold standard: >`atawAo ’ty faEala ’ataw +Verb +Perf +Act +3P +Pl`. This entry has more morphemes than symbols. Such a decomposition is out of the reach of the ANA-SEG model which can only *segment* words. Nevertheless, the results for the ANA-SEG system are rather good for the English considering the simplicity of this approach. In Finnish, however, only about half of the words have been analyzed, which severely impacts recall.

Regarding the results of the ANA-PAIR system, we observe an almost perfect precision but a low recall. This is due to the lack of generalization of the connections. The detected connection are mainly between word sharing the same stem but it does not link together words that share the same affix, which are the main source of morphological relationship. The following entry for the word `phantom’s`: `phantom’s phantom’s+phantoms phantom+phantom’s` illustrates the problem. The link between words sharing the stem `phantom` has been made but it doesn’t link to the thousand of words that share the affix `’s`. Again, such a system was not adapted for the task of the Morpho Challenge but its high precision indicates that formal analogy is a reliable way to related words which might potentially be used by the cofactor-based systems during the construction of the stem tree (Section 3.1.2).

Regarding the cofactor-based systems, we note that the COF-GRAPH model outperforms the COF-FIRST one for all the languages but the Arabic one. The gains in f-score are modest for the English language, but significant for the Turkish language with an absolute improvement of more than 10 points. This performance disparity between languages can be explained by the morphological complexity of the Turkish language. As a matter of fact, in English, applying one `c-rule` to a word is in most cases enough to reach a related word. This is corroborated with the

	COF-GRAPH			COF-FIRST			ANA-SEG			ANA-PAIR		
	Pr.	Rc.	F1	Pr.	Rc.	F1	Pr.	Rc.	F1	Pr.	Rc.	F1
ENG	**78.18**	**47.39**	**59.01**	73.31	47.29	57.49	56.4	32.4	41.2	100	16.2	27.8
FIN	**69.84**	**25.97**	**37.65**	74.82	19.81	31.33	53.2	7.7	13.4	100	0.4	0.8
TUR	**54.60**	**45.36**	**49.55**	47.76	32.92	38.98	58.0	12.4	20.5	100	5.7	10.8
GER	**51.90**	**33.03**	**40.36**	43.93	29.52	35.31	36.4	13.8	20.0	100	3.9	7.4
NVARA	96.04	1.97	3.87	90.38	2.84	5.52	**90.0**	**4.3**	**8.2**	100	2.1	4.1
VOWARA	88.38	1.50	2.95	89.56	2.24	4.37	**86.8**	**2.6**	**5.0**	97.4	1.8	3.6

Table 3: Precision (Pr.), Recall (Rc.) and F-measure (F1) of our systems on MC-REF. Results in bold mark the best results obtained per language.

	ENG			GER		
system	Pr.	Rc.	F1	Pr.	Rc.	F1
COF-GRAPH	**66.11**	**45.95**	**54.01**	**70.35**	**34.38**	**45.70**
COF-FIRST	63.42	45.21	52.54	70.79	30.02	41.69
ANA-SEG	61.89	31.81	41.70	65.27	15.74	24.73
ANA-PAIR	70.64	25.57	37.30	85.21	8.24	14.81

Table 4: Performance of our systems on CELEX-REF.

observation that on CELEX, there is an average of 2.15 morphemes per word (see Table 2). For the Turkish language however, applying more than one `c-rule` gives a clear advantage as it allows to link words with more than two morphemes to its stem, even though no intermediate word exists.

Another gain of using more `c-rules` is the finer grained analysis that results. By applying only one `c-rule` to a word with more than one prefix or suffix, it creates a strong bias toward considering all prefixes or suffixes of this word as a single morpheme. This phenomenon does not impact languages where words often contain only one affix, but has a major impact on more complex languages. Those two phenomenon explain the increase in recall. The gain in precision observed for most of the languages might be explained by the propensity of COF-GRAPH to choose analysis that are supported by many different `c-rules` in the graph.

4.3 Systems tested at Morpho Challenge

At the end, based on the observations we made while developing our different systems, we decided to submit two runs to the shared-task: COF-GRAPH which we named RALI-COF and ANA-SEG which we called RALI-ANA. We did not submit a different run for shared-task 2 and due to time constraints, we decided not to participate to the third task.

5 Discussion and future work

The use of formal analogy for unsupervised morphological analysis is rather new. COF-GRAPH and ANA-SEG are our first systems to participate in the Morpho Challenge competition. Our primary concern was to demonstrate the viability of the approach instead of paying attention to the peculiarities of the shared task. In particular, we did not adjust the meta-parameters controlling our systems for each language.

The development of our systems at a very short notice without adapting them to a language in particular, the simplicity of the models we tested and the decent performance we observed are all facts that reinforce our belief that formal analogy can be used in a morphological analysis task.

The computational issue remains the main obstacle for the deployment of pure analogical-based systems. However, we showed that using cofactors can significantly reduce computation time while providing the best results for all but the Arabic language. Even if in our case, cofactors were still extracted from a large amount of analogies, we think that reliable results could be obtained from smaller lexicons. Preliminary experiments showed that in English, formal analogies computed on less than 10% of the words in the lexicon could identify most of the major affixes.

Our approach on the pure analogical system was quite simple and could gain from using some information theory metric such as perplexity to calculate the probability of the different segmentations. It would be interesting to use the word-context in order to restrain the analogies to words that share some semantic properties, in the vain of the work of Hathout [6].

While `c-rule`s capture more context than cofactors, and therefore reduces some source of noise when using them, we fully realize that other levels of abstraction might be more effective in generalizing the information captured by an analogy. This is left as a future work.

References

[1] R. H. Baayen, R. Piepenbrock, and L. Gulikers. The CELEX lexical database (release 2). CD-ROM, Linguistic Data Consortium, Univ. of Pennsylvania, USA, 1995.

[2] Delphine Bernhard. Simple morpheme labelling in unsupervised morpheme analysis. In *Workshop of Morpho Challenge 2007*, pages 873–880, Budapest, Hungary, Sept. 2007.

[3] Mathias Creutz and Krista Lagus. Inducing the morphological lexicon of a natural language from unannotated text. In *In Proceedings of the International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning (AKRR'05*, volume 5, pages 106–113, 2005.

[4] Étienne Denoual. Analogical translation of unknown words in a statistical machine translation framework. In *Machine Translation Summit, XI*, Copenhagen, Sept. 10-14 2007.

[5] Zellig S. Harris. From phoneme to morpheme. *Language*, 31(2):190–222, 1955.

[6] Nabil Hathout. From wordnet to celex: acquiring morphological links from dictionaries of synonyms. In *Third International Conference on Language Resources and Evaluation*, pages 1478–1484, Las Palmas de Gran Canaria, 2002.

[7] Nabil Hathout. Acquistion of the morphological structure of the lexicon based on lexical similarity and formal analogy. In *3rd Textgraphs workshop on Graph-Based Algorithms in Natural Language Processing*, pages 1–8, Manchester, United Kingdom, Aug. 2008.

[8] Philippe Langlais and Alexandre Patry. Enrichissement d'un lexique bilingue par apprentissage analogique. *Traitement Automatique des Langues (TAL)*, 49 (varia):13–40, 2008.

[9] Philippe Langlais, François Yvon, and Pierre Zweigenbaum. Improvements in analogical learning: Application to translating multi-terms of the medical domain. In *12th Conference of the European Chapter of the Association for Computational Linguistics (EACL-09)*, Athens, Greece, 2009.

[10] Yves Lepage and Étienne Denoual. Purest ever example-based machine translation: Detailed presentation and assessment. *Machine Translation*, 29:251–282, 2005.

[11] Christian Monson, Jaime Carbonell, Alon Lavie, and Lori Levin. Paramor: Minimally supervised induction of paradigm structure and morphological analysis. In *Proceedings of Ninth Meeting of the ACL Special Interest Group in Computational Morphology and Phonology*, pages 117–125, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

[12] Nicolas Stroppa. *Définitions et caractérisations de modèles à base d'analogies pour l'apprentissage automatique des langues naturelles.* PhD thesis, ENST, ParisTech, Télécom, Paris, France, Nov. 2005.

[13] Nicolas Stroppa and François Yvon. An analogical learner for morphological analysis. In *9th Conf. on Computational Natural Language Learning (CoNLL)*, pages 120–127, Ann Arbor, MI, June 2005.

[14] François Yvon, Nicolas Stroppa, Arnaud Delhay, and Laurent Miclet. Solving analogical equations on words. Technical Report D005, École Nationale Supérieure des Télécommunications, Paris, France, Jul. 2004.

[15] Daniel Zeman. Using unsupervised paradigm acquisition for prefixes. In *Workshop of Morpho Challenge 2008*, Arhus, Denmark, Sept. 2008.