EMMA: A Novel *E*valuation *M*etric for *M*orphological *A*nalysis

Sebastian Spiegler
Intelligent Systems Group
University of Bristol
spiegler@cs.bris.ac.uk

Christian Monson
Center for Spoken Language Understanding
Oregon Health & Science University
monsonc@csee.ogi.edu

Abstract

We present a novel *E*valuation *M*etric for *M*orphological *A*nalysis (*EMMA*) that is both linguistically appealing and empirically sound. *EMMA* uses a graph-based assignment algorithm, optimized via integer linear programming, to match morphemes of predicted word analyses to the analyses of a morphologically rich answer key. This is necessary especially for unsupervised morphology analysis systems which do not have access to linguistically motivated morpheme labels. Across 3 languages, *EMMA* scores of 14 systems have a substantially greater positive correlation with mean average precision in an information retrieval (IR) task than do scores from the metric currently used by the Morpho Challenge (MC) competition series. We compute *EMMA* and MC metric scores for 93 separate system-language pairs from the 2007, 2008, and 2009 MC competitions, demonstrating that *EMMA* is not susceptible to two types of gaming that have plagued recent MC competitions: Ambiguity Hijacking and Shared Morpheme Padding. The *EMMA* evaluation script is publicly available from http://www.cs.bris.ac.uk/Research/MachineLearning/Morphology/Resources/.

1 Introduction

Words in natural language are constructed from smaller building blocks called *morphemes*. For example, the word *wives* breaks down into an underlying stem, *wife*, together with a *plural* suffix. Analyzing the morphological structure of words is known to benefit a variety of downstream natural language (NL) tasks such as speech recognition (Creutz, 2006; Arısoy et al., 2009), machine translation (Oflazer et al., 2007), and information retrieval (McNamee et al., 2008).

A variety of automatic systems can morphologically analyze words that have been removed from their surrounding context. These systems range from hand-built finite state approaches (Beesley and Karttunen, 2003) to recently proposed algorithms which learn morphological structure in an unsupervised fashion (Kurimo et al., 2007). Since unsupervised systems do not have access to linguistically motivated morpheme labels, they typically produce morphological analyses that are closely related to the written form. Such a system might decompose *wives* as *wiv -es*. Meanwhile, a hand-built system might propose *wife_N +Plural*, or even parse *wives* as a hierarchical feature structure. As morphological analysis systems produce such varied outputs, comparing decompositions from disparate systems is a challenge.

This paper describes *EMMA*, an *E*valuation *M*etric for *M*orphological *A*nalysis that quantitatively measures the quality of a set of morphological analyses in a linguistically adequate, empirically useful, and novel fashion. *EMMA* evaluates analyses that can be represented as a flat set of symbolic features, including hierarchical representations, which can be projected down to a linearized form (Roark and Sproat, 2007).

An automatic metric that discriminates between proposed morphological analyses should

Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010), pages 1029–1037,
Beijing, August 2010

fulfill certain computational and linguistic criteria. Computationally, the metric should:

1. *Correlate* with the performance of real-world NL processing tasks which embed the morphological analyses.
2. Be *Readily Computable*: The metric will only be useful if it is less time consuming and easier to compute than the larger NL task.
3. Be *Robust*: The metric should be difficult to game and should accurately reflect the distribution of predicted and true morphemes.
4. Be *Readily Interpretable*: When possible, the final numeric score should directly identify the strengths and weaknesses of the underlying morphological analysis system.

While accounting for these computational requirements, a morphology metric should still reward accurate models of linguistic structure. In particular, the metric should account for:

1. *Morphophonology*: Applying a morphological rule may alter the surface form of stem or affix. In the word *wives*, /waivz/, a rule of morphophonology voices the stem-final /f/ of *wife*, /waif/, when the plural suffix is added. A metric should penalize for not placing *wives* and *wife* as forms of the same lexeme.
2. *Allomorphy*: A metric should capture the successful grouping of allomorphs. The German plural has several surface allomorphs including *-en* in *Zeiten* (*times*), *-e* in *Hunde* (*dogs*), and *-s* in *Autos* (*cars*). A metric should reward a morphological analysis system that analyzes the different surface forms of the German plural as underlyingly identical.
3. *Syncretism*: In mirror fashion, a metric should reward analyses that distinguish between surface-identical syncretic morphemes: although *derives* and *derivations* both contain an *-s* morpheme, one marks *3rd person singular* and the other *plural*.
4. *Ambiguity*: Finally, a metric should account for legitimate morphological ambiguity. In Hebrew, the written word *MHGR* has three viable morphological segmentations: *M- H- GR*, "*from the foreigner*", *M- HGR*, "*from Hagar*", and the unsegmented form *MHGR*, meaning "*immigrant*" (Lavie et al., 2004). Absent disambiguating context, a morphological system should be rewarded for calling out all three analyses for *MHGR*.

Morphophonology, allomorphy, syncretism, and ambiguity are all common phenomena in the world's languages. The first three have all received much discussion in theoretical linguistics (Spencer and Zwicky, 2001), while morphological ambiguity has significant practical implications in NL processing, e.g. in machine translation of morphologically complex languages (Lavie et al., 2004; Oflazer et al., 2007).

In Section 2 we propose the metric *EMMA*, which has been specifically designed to evaluate morphological analyses according to our computational and linguistic criteria. Section 3 then describes and qualitatively critiques several well-used alternative metrics. Section 4 empirically compares *EMMA* against the qualitatively-strong metric used in the Morpho Challenge competition series (Kurimo et al., 2009). And we conclude in Section 5.

2 *EMMA*: An *E*valuation *M*etric for *M*orphological *A*nalysis

EMMA, the metric we propose for the evaluation of morphological analyses, like all the metrics that we consider in this paper, compares proposed morphological analyses against an answer key of definitively-analyzed words from a vocabulary. Since a set of proposed analyses is likely to use a different labeling scheme than the answer key, especially true of the output from unsupervised systems, *EMMA* does not perform a direct comparison among proposed and answer analyses. Instead, *EMMA* seeks a one-to-one relabeling of the proposed morphemes that renders them as similar as possible to the answer key. *EMMA*, then, measures the degree to which proposed analyses approximate an isomorphism of the answer key analyses. For exposition, we initially assume that, for each word, a single proposed analysis is scored against a single unambiguous answer analysis. We relax this restriction in Section 2.3, where EMMA scores multiple proposed analyses

against a set of legitimately ambiguous morphological analyses.

To find the most appropriate one-to-one morpheme relabeling, *EMMA* turns to a standard algorithm from graph theory: optimal maximum matching in a bipartite graph. A *bipartite graph*, $G = \{X, Y; E\}$, consists of two disjoint sets of vertices, $X = \{x_1, x_2, \ldots, x_n\}$ and $Y = \{y_1, y_2, \ldots, y_m\}$, and a set of edges $e(x_i, y_j) \in E$ such that each edge has one end in X and the other end in Y. In *EMMA*, the set, A, of all unique morphemes in the answer key and the set, P, of all unique morphemes in the proposed analyses serve as the disjoint vertex sets of a bipartite graph.

A *matching* $M \subseteq E$ in a bipartite graph is defined as a set of edges $e(x_i, y_j)$ such that no x_i or y_j is repeated. A *maximum matching* is a matching where no M' with $|M'| > |M|$ exists. Furthermore, a weight $w(x_i, y_j) \in \Re$ may be assigned to each edge $e(x_i, y_j)$ of a bipartite graph. An *optimal assignment* is a maximum matching which also maximizes the sum of the weights of the edges of the matching

$$\sum_{e(x_i,y_j)\in M} w(x_i, y_j) \ .$$

EMMA weights the edge between a particular answer morpheme $a \in A$ and a proposed morpheme $p \in P$ as the number of words, w, in the vocabulary, V, where the answer analysis of w includes morpheme a while the proposed analysis includes p. *EMMA* constructs an optimal assignment maximum matching in this weighted bipartite morpheme graph. The edge weights ensure that the optimal matching will link the answer and proposed morphemes which globally occur in the analyses of the same words most often – restricting each answer morpheme to be represented by at most one proposed morpheme, and each proposed morpheme to represent at most one morpheme in the answer key. On the one hand, the restrictions thus imposed by bipartite matching penalize sets of proposed analyses that do not differentiate between surface-identical *syncretic morphemes*. On the other hand, the same one-to-one matching restrictions penalize proposed analyses that do not conflate *allomorphs* of the same underlying morpheme, whether those allomorphs are phonologically induced or not. Thus, *EMMA* meets our linguistic criteria from Section 1 of modeling syncretism, allomorphy, and morphophonology.

2.1 Maximum Matching by Integer Linear Programming

To construct the maximum matching optimal assignment of answer and proposed morphemes, *EMMA* uses standard integer linear programming techniques as implemented in *lpsolve* (Berkelaar et al., 2004). For the purpose of our integer program, we represent the weight of each potential edge of the optimal bipartite morpheme assignment in a count matrix $C = \{c_{ij}\}$ where c_{ij} is assigned the number of words $w \in V$ which share morpheme a_i in the answer key and p_j in the prediction. We then define a binary matrix $B = \{b_{ij}\}$ of the same dimensions as C. Each b_{ij} will be set to 1 if an edge exists from a_i to p_j in the optimal maximum matching, with $b_{ij} = 0$ otherwise. The integer linear program can then be defined as follows:

$$\underset{B}{\operatorname{argmax}} \sum_{i,j} (C \cdot B)_{ij} \tag{1}$$

$$s.t. \sum_i b_{ij} \leq 1 \ , \quad \sum_j b_{ij} \ \leq 1 \ , \quad b_{ij} \geq 0 \ ,$$

where $(C \cdot B)_{ij} = c_{ij} \cdot b_{ij}$ is the element-wise *Hadamard product*.

2.2 Performance Measures

Having settled on a maximum matching optimal assignment of proposed and answer morphemes, *EMMA* derives a final numeric score. Let w_k be the k^{th} word of V; and let A_k and P_k denote, respectively, the sets of morphemes in the answer key analysis of w_k and predicted analysis of w_k. Furthermore, let P_k^* denote the predicted morphemes for w_k where a morpheme p_j is replaced by a_i if $b_{ij} = 1$. Now that A_k and P_k^* contain morpheme labels that are directly comparable, we can define *precision* and *recall* scores for the proposed analysis of the word w_k. Precision is the fraction of correctly relabeled proposed morphemes from among all proposed morphemes of w_k; while *recall* is the number of correctly relabeled morphemes as a fraction of the answer key

analysis of w_k. Precision and recall of the full vocabulary are the average word-level precision and recall:

$$precision = \frac{1}{|V|} \sum_{k}^{|V|} \frac{|A_k \bigcap P_k^*|}{|P_k^*|} \quad , \qquad (2)$$

$$recall = \frac{1}{|V|} \sum_{k}^{|V|} \frac{|A_k \bigcap P_k^*|}{|A_k|} \quad . \qquad (3)$$

Finally, *f-measure* is the harmonic mean of precision and recall:

$$f\text{-}measure = \frac{2 \cdot precision \cdot recall}{precision + recall} \quad . \qquad (4)$$

2.3 Morphological Ambiguity in *EMMA*

Thus far we have presented *EMMA* for the scenario where each word has a single morphological analysis. But, as we saw in Section 1 with the Hebrew word *MHGR*, natural language permits surface forms to have multiple legitimate morphological analyses. When a word is truly ambiguous, *EMMA* expects an answer key to contain a set of analyses for that word. Similarly, we permit sets of proposed alternative analyses. To extend *EMMA* with the ability to evaluate alternative analyses we first generalize the optimal maximum matching of morphemes from Section 2.1. We then define a new integer linear program to match answer and proposed *alternative analyses*. Finally, we adjust the performance measures of Section 2.2 to account for alternatives.

2.3.1 Ambiguity and Morpheme Matching

Let $A_{k,r}$ denote the r^{th} alternative answer analysis of the k^{th} word with $1 \leq r \leq m_k$, and let $P_{k,s}$ denote the s^{th} alternative prediction with $1 \leq s \leq n_k$, where m_k is the number of alternative analyses in the answer key and n_k the number of alternative predictions for w_k. We redefine $A_k = \bigcup_r^{m_k} A_{k,r}$ and $P_k = \bigcup_s^{n_k} P_{k,s}$ as the set of all answer or, respectively, predicted morphemes of w_k across all analysis alternatives. Instead of incrementing each c_{ij} entry in the count matrix C by a full count, we now add $\frac{1}{m_k \cdot n_k}$ to c_{ij} for all pairs $(a_i, p_j) \in A_k \times P_k$. This corresponds to counting each combination of an answer key and predicted morpheme normalized by the number of possible pairings between proposed and answer analysis alternatives. When both the answer and proposed analyses consist of just a single alternative, c_{ij} remains unchanged. Generalized morpheme matching still employs the linear program defined in Equation 1.

2.3.2 Matching of Alternative Analyses

After performing a one-to-one morpheme relabelling that accounts for ambiguity, we need to extend *EMMA* with the ability to evaluate alternative analyses. We again turn to optimal maximum matching in a bipartite graph: Where earlier we matched proposed and answer morphemes, now we match full proposed and answer analysis alternatives, maximizing the total number of correctly predicted morphemes across all alternatives. Generalizing on the notation of the unambiguous case, let $P_{k,s}^*$ denote the s^{th} alternative predicted analysis of the k^{th} word where predicted morphemes have been replaced by their assigned answer key morphemes. We introduce a new count matrix $C' = \{c'_{r,s}\}$, where $c'_{r,s}$ is the count of common morphemes of the r^{th} answer key alternative and s^{th} predicted alternative. Based on Equation 1, we calculate the binary matrix $B' = \{b'_{r,s}\}$ which contains the optimal assignment of the alternative answer key and predicted analyses for w_k.

2.3.3 Ambiguity and Performance Scores

We now adjust *EMMA*'s numeric performance measures to account for sets of ambiguous analysis alternatives. *Precision* becomes

$$\frac{1}{|V|} \sum_{k}^{|V|} \frac{1}{n_k} \sum_{r}^{m_k} \sum_{s}^{n_k} b'_{r,s} \frac{|A_{k,r} \bigcap P_{k,s}^*|}{|P_{k,s}^*|} \quad , \qquad (5)$$

the ratio of correctly predicted morphemes across all predicted alternatives normalised by the number of predicted alternatives, n_k, and the vocabulary size, $|V|$. The factor $b'_{r,s}$ guarantees that scores are only averaged over pairs of proposed and answer analysis alternatives that have been assigned, that is, where $b'_{r,s} = 1$. *Recall* is measured similarly with

$$\frac{1}{|V|} \sum_{k}^{|V|} \frac{1}{m_k} \sum_{r}^{m_k} \sum_{s}^{n_k} b'_{r,s} \frac{|A_{k,r} \bigcap P_{k,s}^*|}{|A_{k,r}|} \quad . \qquad (6)$$

Here, we normalize by m_k, the number of alternative analyses for the k^{th} word that are listed in the answer key. The normalisation factors $\frac{1}{m_k}$ and $\frac{1}{n_k}$ ensure that predicting too few or many alternative analyses is penalised.

3 Other Morphology Metrics

Having presented the *EMMA* metric for evaluating the quality of a set of morphological analyses, we take a step back and examine other metrics that have been proposed. Morphology analysis metrics can be categorized as either: 1. Directly comparing proposed analyses against an answer key, or 2. Indirectly comparing proposed and answer analyses by measuring the strength of an isomorphic-like relationship between the proposed and answer morphemes. The proposed *EMMA* metric belongs to the second category of isomorphism-based metrics.

3.1 Metrics of Direct Inspection

By Segmentation Point. Perhaps the most readily accessible automatic evaluation metric is a direct comparison of the morpheme boundary positions in proposed and answer analyses. As early as 1974, Hafer and Weiss used the direct boundary metric. Although intuitively simple, the segmentation point method implicitly assumes that it is possible to arrive at a valid morphological analysis by merely dividing the characters of a word into letter sequences that can be reconcatenated to form the original word. But, by definition, concatenation cannot describe non-contanative processes like morphophonology and allomorphy. Nor does simple segmentation adequately differentiate between surface-identical syncretic morphemes. Despite these drawbacks, precision and recall of segmentation points is still used in current morphological analysis research (Poon et al. (2009), Snyder and Barzilay (2008), Kurimo et al. (2006)).

Against Full Analyses. To confront the reality of non-concatenative morphological processes, an answer key can hold full morphological analyses (as opposed to merely segmented surface forms). But while a hand-built (Beesley and Karttunen, 2003) or supervised (Wicentowski , 2002) morphology analysis system can directly model the annotation standards of a particular morphological answer key, the label given to specific morphemes is ultimately an arbitrary choice that an unsupervised morphology induction system has no way to discover.

By Hand. On the surface, scoring proposed analyses by hand appears to provide a way to evaluate the output of an unsupervised morphology analysis system. Hand evaluation, however, does not meet our criteria from Section 1 for a robust and readily computable metric. It is time consuming and, as Goldsmith (2001) explains, leaves difficult decisions of what constitutes a morpheme to on-the-fly subjective opinion.

3.2 Metrics of Isomorphic Analysis

Recognizing the drawbacks of direct evaluation, Schone and Jurafsky (2001), Snover et al. (2002), and Kurimo et al. (2007) propose related measures of morphological analysis quality that are based on the idea of an isomorphism. For reasons that will be clear momentarily, we refer to the Schone and Jurafsky, Snover et al., and Kurimo et al. metrics as *soft* isomorphic measures. As discussed in Section 2, metrics of isomorphism measure similarities between the distribution of proposed morphemes and the distribution of answer morphemes, where proposed and answer morphemes may be disjoint symbol sets.

Unlike the *EMMA* metric proposed in Section 2, the soft metrics of isomorphism do not seek to explicitly link proposed morphemes to answer morphemes. Instead, their metrics group sets or pairs of words which share, in *either* the proposed analyses or in the answer analyses, a stem (Schone and Jurafsky, 2001; Snover, 2002), a suffix (Snover et al., 2002), or any arbitrary morpheme (Kurimo et al., 2007). The soft metrics subsequently note whether these same sets or pairs of words share any morpheme in the answer key or, respectively, in the proposed analyses. By foregoing a hard morpheme assignment, the soft metrics do not adequately punish sets of proposed and answer morphemes which fail to model syncretism and/or allomorphy. For example, proposed analyses that annotate *3rd person singular* and *plural* with a single undifferentiated *+s* morpheme will receive recall credit for both nouns and

verbs.

3.3 The Morpho Challenge Metric

The Morpho Challenge (MC) competition series for unsupervised morphology analysis algorithms (Kurimo et al., 2009) has used a soft metric of isomorphism in its most recent three years of competition: 2007, 2008, and 2009. According to Kurimo et al. (2009) the Morpho Challenge (MC) measure samples *random word pairs* which share at least one common morpheme. Precision is calculated by generating random word pairs from the set of proposed analyses and then comparing the analyses of the word pairs in the answer key. The fraction of found and expected common morphemes is normalised by the number of words which are evaluated. *Recall* is defined in mirror fashion. The MC metric also normalizes precision and recall scores across sets of alternative analyses for each word in the proposal and answer key. To our knowledge the MC metric is the first isomorphism-based metric to attempt to account for *morphological ambiguity*. As we show in Section 4, however, MC's handling of ambiguity is easily gamed.

The MC metric does meet our criterion of being *readily computable* and, as we will show in the experimental section, the metric also *correlates* to a certain extent with performance on a higher-level natural language processing task. The downside of the MC metric, however, is *robustness*. In addition to MC's crude handling of ambiguity and its over-counting of allomorphs and syncretic morphemes, the random pair sampling method that MC uses is not independent of the set of analyses being evaluated. If two algorithms predict different morpheme distributions, the sampling method will find different numbers of word pairs. We substantiate our claim that the MC metric lacks robustness in Section 4 where we empirically compare it to the *EMMA* metric.

4 Experimental Evaluation

To experimentally evaluate our newly proposed *EMMA* metric, and to quantitatively compare the *EMMA* and MC metrics, we have evaluated results of 93 system-language pairs from Morpho Challenge 2007, 2008, and 2009.[1] The evaluation comprised three algorithms by Bernhard (2007) and Bernhard (2009), one algorithm by Can and Manandhar (2009), the MC baseline algorithm *Morfessor* by Creutz (2006), *UNGRADE* by Golenia et al. (2009), two algorithms by Lavallee and Langlais (2009), one algorithm by Lignos et al. (2009), five *ParaMor* versions by Monson et al. (2008) and Monson et al. (2009), three *Promodes* versions by Spiegler et al. (2009) and one algorithm by Tchoukalov et al. (2009). We ran these algorithms over six data sets available from the MC competition: Arabic (vowelized and non-vowelized), English, Finnish, German, and Turkish. We then scored the system outputs using both *EMMA* and the MC metric against an answer key provided by MC. In Sections 2 and 3.3 we have already commented on the *linguistic characteristics* of both metrics. In this section, we concentrate on their *computational performance*.

Both the *EMMA* and MC metrics are *readily computable*: Both are freely available[2] and they each take less than two minutes to run on the average desktop machines we have used. In terms of *interpretability*, *EMMA* not only returns the performance as precision, recall and f-measure as MC does, but also provides predicted analyses where mapped morphemes are replaced by answer key morphemes. This information is helpful when judging results qualitatively since it exposes tangible algorithmic characteristics. In Table 1 we present the algorithms with the highest MC and *EMMA* scores for each language. For all languages, the *EMMA* and MC metrics place different algorithms highest. One reason for the significantly different rankings that the two metrics provide may be the *sampling of random pairs* that MC uses. Depending on the distribution of predicted morphemes across words, the number of random pairs, which is used for calculating the precision, may vary. For instance, on vowelized Arabic, *Promodes 1* is evaluated over a sample of 100 pairs where MC selected just 47 pairs for *ParaMor Mimic*.

[1] Detailed results can be found in Spiegler (2010).

[2] *EMMA* may be downloaded from `http://www.cs.bris.ac.uk/Research/MachineLearning/Morphology/Resources/`

Language	Algorithm and year of participation in MC		MC evaluation metric			*EMMA* evaluation metric		
			Pr.	Re.	F1	Pr.	Re.	F1
Arabic (nv)	Promodes 2	2009	0.7789	0.3980	**0.5268**	0.5356	0.2444	0.3356
	Ungrade	2009	0.7971	0.1603	0.2670	0.7017	0.2490	**0.3675**
Arabic (vw)	Promodes 2	2009	0.5946	0.6017	**0.5982**	0.4051	0.3199	0.3575
	Promodes 1	2009	0.7381	0.3477	0.4727	0.5588	0.3281	**0.4135**
English	Bernhard 1	2007	0.7850	0.5763	**0.6647**	0.8029	0.7460	0.7734
	Lignos	2009	0.7446	0.4716	0.5775	0.9146	0.6747	**0.7766**
Finnish	ParaMorPlusMorfessor	2008	0.5928	0.5675	**0.5798**	0.2271	0.3428	0.2732
	Lavallee rali-cof	2009	0.6731	0.3563	0.4659	0.5061	0.4065	**0.4509**
German	ParaMorPlusMorfessor	2008	0.5562	0.6077	**0.5808**	0.3633	0.4948	0.4190
	Morfessor	2009	0.6528	0.3818	0.4818	0.7311	0.5556	**0.6314**
Turkish	ParaMorPlusMorfessor	2008	0.6779	0.5732	**0.6212**	0.3476	0.4315	0.3851
	Morfessor	2009	0.7894	0.3330	0.4684	0.5901	0.3703	**0.4550**

Table 1: Best performing algorithms with MC and *EMMA* evaluation metric.

Algorithm and year of participation in MC		MC evaluation metric			*EMMA* evaluation metric		
		Pr.	Re.	F1	Pr.	Re.	F1
Morfessor	2009	0.8143	0.2788	0.4154	0.4751	0.3472	0.4012
ParaMor	2008	0.4111	0.4337	0.4221	0.4322	0.3770	0.4027
ParaMorPlusMorfessor	2008	0.5928	0.5675	**0.5798**	0.2271	0.3428	**0.2732**
Paramor Morfessor Union	2009	0.4374	0.5676	0.4941	0.3878	0.4530	0.4178

Table 3: Gaming MC with ambiguity hijacking on Finnish.

Looking at any particular algorithm-language pair, the *EMMA* and MC scores differ considerably and respective raw scores are not directly comparable. More interesting is the extent to which both metrics *correlate* with real NL tasks. Table 2 lists the Spearman rank correlation coefficient for algorithms from MC 2009 on English, Finnish and German comparing rankings of f-measure results returned by either MC or *EMMA* against rankings using the mean average precision (MAP) of an information retrieval (IR) task.[3] All MAP scores are taken from Kurimo et al. (2009). Although both metrics positively correlate with the IR results; *EMMA*'s correlation is clearly stronger across all three languages.

To test the *robustness* of the *EMMA* and MC metrics, we performed two experiments where we intentionally attempt to game the metrics – *ambiguity hijacking* and *shared morpheme padding*. In both experiments, the MC metric showed vulnerability. Ambiguity hijacking results for Finnish appear in Table 3, other languages perform similarly. Using both metrics, we scored the Finnish analyses that were proposed by a) the *Morfessor* algorithm alone, b) *ParaMor* alone, and c) two ways of combining ParaMor and Morfessor: *ParaMorPlusMorfessor* simply lists the ParaMor and Morfessor analyses as alternatives – as if each word were ambiguous between a ParaMor and a Morfessor analysis; *ParaMorMorfessorUnion*, on the other hand, combines the morpheme boundary predictions of *ParaMor* and *Morfessor* into a single analysis. The *ParaMorPlusMorfessor* system games the ambiguity mechanism of the MC metric, achieving an f-measure higher than that of any of the three other algorithms. *EMMA*, however, correctly discovers that the analyses proposed by *ParaMorPlusMorfessor* lie farther from an isomorphism to the the answer key than do the unified analyses of *ParaMorMorfessorUnion*.

In Table 4 we show a second way of gaming the MC metric – *shared morpheme padding*. We add the same unique bogus morpheme to each proposed analysis of every word for all systems.

[3] Detailed results can be found in Spiegler (2010).

Language	MC evaluation			EMMA evaluation		
	Precision	Recall	F-measure	Precision	Recall	F-measure
Arabic (nv)	0.91±0.02	**10.83 ± 8.33**	**7.20±5.10**	0.91±0.05	1.30±0.07	1.20±0.05
Arabic (vw)	0.85±0.04	**11.17±8.81**	**7.13±5.23**	0.89±0.07	1.21±0.06	1.12±0.05
English	**0.36±0.08**	**2.02±0.66**	0.63±0.10	0.73±0.15	1.05±0.08	0.86±0.12
Finnish	0.57±0.08	**3.07±2.47**	1.19±0.68	0.87±0.19	1.12±0.10	0.99±0.14
German	**0.43±0.08**	**2.90±1.45**	0.84±0.16	0.80±0.17	1.09±0.08	0.94±0.11
Turkish	0.58±0.09	**2.95±1.65**	1.19±0.37	0.85±0.08	1.07±0.04	0.97±0.05

Table 4: Gaming MC with shared morpheme padding: Average and standard deviations of the ratio of padded to original scores.

Padding analyses with a shared morpheme significantly increases the recall scores of the MC metric. We summarize our experimental results by calculating, for each language-algorithm pair, the ratio of the score for the padded analyses as compared to that of the original, unpadded analyses. Table 4 reports average and standard deviation of the ratios across all systems for each language. In Arabic (nv. and vw.), the recall increases by 10.83 and 11.17 times, which leads to an inflation of f-measure by 7.20 and 7.13 times – this is a direct result of the soft nature of the MC isomorphism. In contrast, *EMMA*'s recall scores increase much less than MC's do, and *EMMA*'s precision scores decrease proportionately. A small change to the set of proposed analyses does not lead to a huge difference in f-measure – characteristic of a more *robust* metric.

5 Conclusion

This paper has proposed, *EMMA*, a novel evaluation metric for the assessment of the quality of a set of morphological analyses. *EMMA*'s:

1. Coverage of the major morphological phenomena,
2. Correlation with performance on natural language processing tasks, and
3. Computational robustness

all recommend the the metric as a strong and useful measure – particularly when evaluating unsupervised morphology analysis systems which, lacking access to labeled training data, are uninformed of the labeling standard used in the answer key.

	Correlation with IR	
	IR vs. MC	IR vs. *EMMA*
English	0.466	**0.608**
Finnish	0.681	**0.759**
German	0.379	**0.637**

Table 2: Spearman rank correlation coefficient of metrics vs. Information Retrieval (IR).

Acknowledgements

We would like to acknowledge various fruitful discussions with Aram Harrow, Alex Popa, Tilo Burghardt and Peter Flach. The work was partially sponsored by EPSRC grant EP/E010857/1 *Learning the morphology of complex synthetic languages*, as well as by NSF Grant #IIS-0811745 and DOD/NGIA grant #HM1582-08-1-0038.

References

Arısoy, Ebru, Doğan Can, Sıddıka Parlak, Haşim Sak, and Murat Saraçlar. 2009. Turkish Broadcast News Transcription and Retrieval. *IEEE Trans. on Audio, Speech and Lang. Proc.*

Beesley, Kenneth R. and Lauri Karttunen. 2003. *Finite State Morphology*. University of Chicago Press.

Berkelaar, Michel, Kjell Eikland, and Peter Notebaert. 2004. Open source (mixed-integer) linear programming system, version 5.1.0.0. `http://lpsolve.sourceforge.net/`.

Bernhard, Delphine. 2007. Simple morpheme labelling in unsupervised morpheme analysis. *Working Notes, CLEF 2007 Workshop*.

Bernhard, Delphine. 2009. Morphonet: Exploring the use of community structure for unsupervised morpheme analysis. *Working Notes, CLEF 2009 Workshop*.

Can, Burcu and Suresh Manandhar. 2009. Unsupervised learning of morphology by using syntactic categories. *Working Notes, CLEF 2009 Workshop*.

Creutz, Mathias. 2006. *Induction of the Morphology of Natural Language: Unsupervised Morpheme Segmentation with Application to Automatic Speech Recognition*. Ph.D. thesis, Helsinki University of Technology, Espoo, Finland.

Goldsmith, John. 2001. Unsupervised learning of the morphology of a natural language. *Comp. Ling.*, 27.

Golénia, Bruno, Sebastian Spiegler, and Peter Flach. 2009. Ungrade: unsupervised graph decomposition. *Working Notes, CLEF 2009 Workshop*.

Hafer, M. A. and S. F. Weiss. 1974. Word segmentation by letter successor varieties. *Inf. Storage and Retrieval*, 10.

Kurimo, Mikko, Mathias Creutz, Matti Varjokallio, Ebru Arisoy, Murat Saraclar. 2006. Unsupervised segmentation of words into morphemes - Morpho Challenge 2005. *Interspeech*.

Kurimo, Mikko, Mathias Creutz, and Ville Turunen. 2007. Overview of morpho challenge in CLEF 2007. *Working Notes, CLEF 2007 Workshop*.

Kurimo, Mikko and Ville Turunen. 2008. Unsupervised Morpheme Analysis Evaluation by IR experiments – Morpho Challenge 2008. *Working Notes, CLEF 2008 Workshop*.

Kurimo, Mikko, Sami Virpioja, and Ville T. Turunen. 2009. Overview and results of morpho challenge 2009. *Working Notes, CLEF 2009 Workshop*.

Lavallee, Jean-Francois and Philippe Langlais. 2009. Morphological Acquisition by Formal Analogy. *Working Notes, CLEF 2009 Workshop*.

Lavie, Alon, Erik Peterson, Katharina Probst, Shuly Wintner, Yaniv Eytani. 2004. Rapid Prototyping of a Transfer-based Hebrew-to-English Machine Translation System. *Proc. of TMI-2004*.

Lignos, Constantine, Erwin Chan, Mitchell P. Marcus, and Charles Yang. 2009. A rule-based unsupervised morphology learning framework. *Working Notes, CLEF 2009 Workshop*.

McNamee, Paul, Charles Nicholas, and James Mayfield. 2008. Don't Have a Stemmer? Be Un+concern+ed *Proc. of the 31st Anual International ACM SIGIR Conference 20-24 July 2008*.

Monson, Christian, Jaime Carbonell, Alon Lavie, and Lori Levin. 2008. Paramor and morpho challenge 2008. *Working Notes, CLEF 2008 Workshop*.

Monson, Christian, Kristy Hollingshead, and Brian Roark. 2009. Probabilistic paramor. *Working Notes, CLEF 2009 Workshop*.

Oflazer, Kemal, and İlknur Durgar El-Kahlout. 2007. Different Representational Units in English-to-Turkish Statistical Machine Translation. *Proc. of Statistical Machine Translation Workshop at ACL 2007*.

Poon, Hoifung, Colin Cherry and Kristina Toutanova 2009. Unsupervised Morphological Segmentation with Log-Linear Models. *Proc. of ACL*.

Roark, Brian and Richard Sproat. 2007. *Computational Approaches to Morphology and Syntax*. Oxford Univ. Press.

Schone, Patrick and Daniel Jurafsky. 2001. Knowlege-free induction of inflectional morphologies. *Proc. of NAACL-2001*.

Snover, Matthew G., Gaja E. Jarosz and Michael R. Brent. 2002. Unsupervised Learning of Morphology Using a Novel Directed Search Algorithm: Taking the First Step. *Proc. of the ACL-02 SIGPHON Workshop*.

Snyder, Benjamin and Regina Barzilay. 2008. Unsupervised Multilingual Learning for Morphological Segmentation. *Proc. of ACL-08: HLT*.

Spencer, Andrew and Arnold M. Zwicky, editors. 2001. *The Handbook of Morphology*. Wiley-Blackwell.

Spiegler, Sebastian, Bruno Golénia, and Peter A. Flach. 2009. Promodes: A probabilistic generative model for word decomposition. *Working Notes, CLEF 2009 Workshop*.

Spiegler, Sebastian. 2010. *EMMA: A Novel Metric for Morphological Analysis - Experimental Results in Detail*. Computer Science Department, University of Bristol, U.K.

Tchoukalov, Tzvetan, Christian Monson, and Brian Roark. 2009. Multiple sequence alignment for morphology induction. *Working Notes, CLEF 2009 Workshop*.

Wicentowski, Richard 2002. *Modeling and Learning Multilingual Inflectional Morphology in a Minimally Supervised Framework*. Ph.D. thesis, The Johns Hopkins University, Baltimore, Maryland, U.S.A.