Subword Variation in Text Message Classification

Robert Munro
Department of Linguistics
Stanford University
Stanford, CA 94305
rmunro@stanford.edu

Christopher D. Manning
Department of Computer Science
Stanford University
Stanford, CA 94305
manning@stanford.edu

Abstract

For millions of people in less resourced regions of the world, text messages (SMS) provide the only regular contact with their doctor. Classifying messages by medical labels supports rapid responses to emergencies, the early identification of epidemics and everyday administration, but challenges include text-brevity, rich morphology, phonological variation, and limited training data. We present a novel system that addresses these, working with a clinic in rural Malawi and texts in the Chichewa language. We show that modeling morphological and phonological variation leads to a substantial average gain of F=0.206 and an error reduction of up to 63.8% for specific labels, relative to a baseline system optimized over word-sequences. By comparison, there is no significant gain when applying the same system to the English translations of the same texts/labels, emphasizing the need for subword modeling in many languages. Language independent morphological models perform as accurately as language specific models, indicating a broad deployment potential.

1 Introduction

The whole world is texting, but rarely in English. Africa has seen the greatest recent uptake of cellphones, with an 8-fold increase over the last 5 years and saturation possible in another 5 (Buys et al., 2009). This is a leapfrog technology – for the majority of new users cellphones are the *only* form of remote communication, surpassing landlines, (non-mobile) internet access and even grid electricity, with costs making texts the dominant communication method. This has led social development organizations to leverage mobile technologies to support health (Leach-Lemens, 2009), banking (Peevers et al., 2008), access to market information (Jagun et al., 2008), literacy (Isbrandt, 2009) and emergency response (Munro, 2010). The possibility to automate many of these services through text-classification is huge, as are the potential benefits – those with the least resources have the most to gain.

However, the data presents many challenges, as text messages are brief, most languages have rich morphology, spellings may be overly-phonetic, and there is often limited training data. We partnered with a medical clinic in rural Malawi and FrontlineSMS:Medic, whose text message management systems serve a patient population of over 2 million in less developed regions of the world. The system allows remote community health workers (CHWs) to communicate directly with more qualified medical staff at centralized clinics, many for the first time.

We present a short-message classification system that incorporates morphological and phonological/orthographic variation, with substantial improvements over a system optimized on word-sequences alone. The average gain is F=0.206 with an error reduction of up to 63.8% for specific labels. For 6 of the 9 labels this more than *doubles* the accuracy. By comparison, there is *not* a significant gain in accuracy when applying the same system to the English translations of the same texts/labels, emphasizing the need for modeling subword structures, but also highlighting why morphology has been peripheral in text classification until now.

Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL, pages 510–518,
Los Angeles, California, June 2010.

2 Language and data

Chichewa is a Bantu language with about 13 million speakers in Southern Africa including 65% of Malawians. We limit examples to the nouns: *odwala* 'patient', *mankhwala* 'medicine'; verb: *fun* 'want'; and the 1st person pronoun/marker: *ndi-* 'I'. Chichewa is closely related to many neighboring languages – more than 100 million people could recognize *ndifuna* as 'I want'.

The morphological complexity is average with about 2-3 morpheme boundaries per word, but this is rich and complex compared to estimates for English, Spanish and Chinese with average of 0.33, 0.85 and 0.01 morpheme boundaries per word. A typical verb is *ndimakafunabe*, 'I am still wanting', consisting of six morphemes, *ndi-ma-ka-fun-a-be*, expressing: 1st person Subject; present tense; noun-class (gender) agreement with the Object; 'want'; verb part-of-speech; and incompletive aspect.

2.1 Labels

The text messages are coded for 0-9 labels in 3 groupings (with counts):

Administrative: related to the clinic:

1. Patient-related (394)
2. Clinic-admin: meetings, supplies etc (169)
3. Technological: phone-credit, batteries etc (21)

Requests: from Community Health Workers:

4. Response: any action requested by CHW (124)
5. Request for doctor (62)
6. Medical advice: CHW asking for advice (23)

Illness: changes of interest to monitoring bodies:

7. TB: tuberculosis (44)
8. HIV: HIV, AIDS and/or treatments (45)
9. Death: reported death of a patient (30)

The groupings correspond to the three main stakeholders of the messages: the clinic itself, interested in classifying messages according to internal work-practices; the Community Health Workers and their patients, acting as the direct care-givers outside the clinic; and broader bodies like the World Health Organization who are interested in monitoring diseases and early identification of epidemics (biosurveillance). The labels are the three most frequent labels required by each of these user groups.

We analyzed 4 months of texts messages with approximately 1,500 labels from 600 messages, consisting of 8,000 words and 30,000 morphemes. While this is small, the final system is being piloted at a clinic in rural Malawi, where users can define new labels at any time according to changing work-practices, new diseases etc. If more than 4 months of manually labeling were required it could limit the utility and user acceptance.

All the messages were translated into English by a medical practitioner, allowing us to make cross-linguistic comparisons of our system.

2.2 Variation

The variation in the data is large. There are >40 forms for 'patient' and only 32% are *odwala*. Of the rest, >50% occur only once. The variation results from morphology: *ndi-odwala*; phonology: *odwa*<u>*r*</u>*a*, *ndi*<u>*w*</u>*odwala*, and compounding: *ndatindidziwewodwala*. There are also >10 spellings for the English borrowing: *patient*, *pachenti* etc, and 3 for the synonym *matenda*.

Similarly, there are >20 forms for 'medicine'. For *fun* 'want', there are >30 forms with >80% occuring only once. There are >200 forms containing *ndi* and no one form accounts for more than 5% of the instances.

The co-occurrence of *ndi* and *fun* within a word is a strong non-redundant predictor for several labels, but >75% of forms occur only once and >85% of the forms are non-contiguous, as above and in the most frequent *ndi-ma-funa* 'I currently want'.

By contrast, in the English translations 'needing' occurs just once but all other forms of 'patient', 'medicine' and '(I) want/need' are frequent.

This brief introduction to the language and data should make it clear that specialized methods are required for modeling variation in text messages, especially in many languages where text messaging is the dominant form of digital communication.

3 Morphological models

We compared language specific and language independent morphological models, comparing 3 methods (with *ndimafuna* as an example):

Stemmed: {*ndi, fun*}
Segmented: {*ndi, ma, fun, a*}
Morph-config: {*ndi-ma, ndi-fun, ndi-a, ma-fun...*}

We also looked at character ngrams, as used by Hidalgo et al. (2006) for morphological variation in English and Spanish. The results converged with those of the segmented model, which is not surprising as the most frequent features would be similar and increasing data items would overcome the sparcity. We leave more sophisticated character ngram modeling for future work.

3.1 Language specific

For the language specific morphological models we implemented a morphological parser as a set of context-free grammars for all possible prefixes and suffixes according to the formal definitions of Chichewa morphology in Mchombo (2004).

We identified stems by parsing potential prefixes and suffixes, segmenting a word w into n morphemes $w_{m,0}, \ldots, w_{m,n-1}$ leaving a stem w_s with length $len(w_s)$ and corpus frequency of $f(w_s)$, such that $len(w_s) > 0$ (ie, there must be a stem). Where multiple parses could be applied, we minimized $len(w_s)$, then maximized n.

3.2 Language independent

For the language independent morphological models we adapted the word-segmenter of Goldwater, Griffiths and Johnson (2009), to morphological parsing (see Related Work for other algorithms we tested/considered). It was suited to our task because a) it is largely nonparametric, meaning that it can be deployed as a black-box before language-specific properties are known b) it favored recall over precision (see the Results for discussion) and c) using a segmentation algorithm, rather than explicitly modeling morphology, also addresses compounds.

This model uses a *Hierarchical Dirichlet Process* (HDP) (Teh et al., 2005). Every morpheme in the corpus m_i is drawn from a distribution G which consists of possible morphemes (the affixes and stems) and probabilities associated with each morpheme. G is generated from a Dirichlet Process (DP) distribution $DP(\alpha_0, P_0)$, with morphemes sampled from P_0 and their probabilities determined by a concentration parameter α_0. The context-sensitive model where H_m is the DP for a specific morpheme is:

$$\begin{array}{lll} m_i | m_{i-1} = m, H_m & \sim H_m & \forall m \\ H_m | \alpha_1, G & \sim DP(\alpha_1, G) & \forall m \\ G | \alpha_0, P & \sim DP(\alpha_0, P_0) & \end{array}$$

Note that this part of our model is identical to the bigram HDP in Goldwater et al. (2009), except that we possess a set of morphemes, not words. Because word boundaries are already marked in the majority of the messages, we constrain the model to treat all existing word boundaries in the corpus as morpheme boundaries, thus constraining the model to morpheme and compound segmentation.

Unlike word-segmentation, not all tokens in the morpheme lexicon are equal, as we want to model stems separately from affixes in the stemmed models. We assume a) the free morphemes (stems and through compounding) are the least frequent and therefore have the lowest final probability, $P(m)$, in the HDP model; and b) each word w must have at least one free morpheme, the stem w_s ($w_s \neq \emptyset$).[1]

The token-optimal process for identifying stems is straightforward and efficient. The words are sorted by the $argmin$ probabilities of $P(w_{m,0}), \ldots, P(w_{m,n-1})$. For each word w, unless w_s can be identified by a previously observed free morpheme, w_s is identified as $argmin(P(w_{m,0}), \ldots, P(w_{m,n-1}))$ and w_s is added to our lexicon of free morphemes. This algorithm iterates over the words with one extra pass to mark all free morphemes in each word (assuming that there might be compounds we missed on the first pass). The cost, where M is the total number of morphemes and W the total number of words, is $O(log(W) + M)$.

This process has the potential to miss free morphemes that only happened to occur in compounds with less-probable stems, but this did not occur in our data.

4 Phonological/Orthographic Models

We compared three models of phonological/orthographic variation:

Chichewa: Chichewa specific
Script: Roman script specific
Indep: language independent

We refer to these using the term ‘phonology’ very broadly. The majority of the variation stems from

[1] Note that identifying stems must be a separate step – if we allowed multiple free morphemes for each word to enter the lexicon without penalty in the HDP model it would converge on a zero-penalty distribution where *all* morphemes were free.

the phonology, but also from phonetic variation as expressed in a given writing system, and variation in the writing system itself arising from fluent speakers with varying literacy.

4.1 Chichewa specific

For the language specific normalization, we applied a set of heuristics to the data, based on the variation given in (Paas, 2005) and our own knowledge of how Bantu languages are expressed in Roman scripts. The heuristics were used to normalize all alternates, eg: $\{iwo \rightarrow i\emptyset o\}$ and $\{r \rightarrow l\}$, resulting in *ndiwodwara* → *ndiodwala*.

The heuristics represented forms for phonemes with the same potential place of articulation ('c/k'), forms with an adjacent place-of-articulation that are common phonological alternates ('l/r', 'e,i'), voicing alternations ('s/z'), or language-internal phonological processes like the insertion of a glide between vowels that the morphology has made adjacent (like we pronounce but don't spell in 'go(w)ing' in English).

We also implemented hard-coded acronym-recovery methods for acronyms associated with the 'Illness' labels: 'HIV', 'TB', 'AIDS', 'ARV'.

4.2 Script specific

The script specific techniques used the same sets of alternates in the language specific model, but normalized such that the heuristic H was applied to a word w in the corpus C resulting in an alternate w', iff $w' \in C$. This method limits the alternates to those whose existence is supported by the data. It is therefore more conservative than the previous method.

For more general acronym identification, we adapted the method of Schwartz & Hearst (2003). We created a set of candidate acronyms by identifying capitalized sequences in non-capitalized contexts and period-delimited single character sequences. All case-insensitive sequences that were segmented by consistent non-alphabetic characters were then identified as acronyms, provided that they ended in a non-alphabetic character. We could not define a similar acronym-start boundary, as prefixes were often added to acronyms, even when the acronyms themselves contained spaces, eg: '*aT. B.*'.

4.3 Language independent

For complete language independence we applied a noise-reduction algorithm to the stream of characters in order to learn the heuristics that represented potential phonological alternates by identifying all minimal pairs of characters sequences (sequences that alternated by one character, include the absence of a character).

Given all sequences of characters, we identified all pairs of sequences of length $> l$ that differed by one character c_1, where c_1 could be null. We then ranked the pairs of alternating sequences by descending length and applied a threshold t, selecting the t longest sequences, creating alternating patterns from all pairs. Regardless of l or t, the resulting heuristics did not resemble those in 4.1 or 4.2.

We did not implement any acronym identification methods, for obvious reasons.

5 Results

The results are compared to a baseline system optimized over word sequences (words and ngrams but no subword modeling). All results presented here are from a MaxEnt model using a leave-one-out cross-validation.

For the English translations of the texts there was no phonological/orthographic variation beyond that resulting from morphology, so we only applied the language independent morphological models.

5.1 Morphology

With the exception of the unsupervised stemming, all the morphological models led to substantial gains in accuracy. As Table 1 shows, the most accurate system used the language specific segmentation, with an average accuracy of F=0.476, a macro-average gain of 22.4%.

The greatest increase in accuracy occured where verbs were the best predictors – the words with the most complex morphology. The 'Response' label showed the greatest relative gain in accuracy for those with a non-zero baseline, where the accuracy increased 4-fold from F=0.113 to F=0.442. It is expected that a label predicated on requests for action should rely on the isolation of verb stems, but this is still a very substantial gain. In contrast to this 391.2% gain in accuracy for Chichewa, the gain for

	Baseline	Stemmed		Segmented		Morph-Config		Gain	
Label		Chich	Indep	Chich	Indep	Chich	Indep	Best	Final
Patient-related	0.830	0.842	0.735	0.857	0.832	0.851	0.867	+3.7	+3.7
Clinic-admin	0.358	0.490	0.295	0.612	0.561	0.577	0.580	+25.5	+22.2
Technological	0	0	0	0.320	0.174	0.320	0.091	+32.0	+09.1
Response	0.113	0.397	0.115	0.440	0.477	0.459	0.442	+36.4	+32.9
Request for doctor	0.121	0.312	0.090	0.505	0.395	0.477	0.375	+38.4	+25.4
Medical advice	0	0	0	0.083	0.160	0.083	0.083	+16.0	+08.3
HIV	0.379	0.597	0	0.554	0.357	0.484	0.351	+21.8	(-2.8)
TB	0.235	0.357	0	0.414	0.200	0.386	0.327	+17.8	+09.2
Death	0.235	0.333	0.229	0.500	0.667	0.462	0.723	+48.8	+48.8
Average.	0.252	0.370	0.163	0.476	0.425	0.455	0.427	+22.4	+17.4

Table 1: Morphology results: F-values for leave-one-out cross-validation comparing different morphological models. *Indep* = language independent, *Chich* = specific to Chichewa, ( ) = not significant ($\rho > 0.05, \chi^2$), *Final* = Gain of the 'Morph-Config, Indep' model over the Baseline.

English, while still relying on the isolation of verb stems, only increased the accuracy by 5.4%.

The unsupervised stemming underperformed the baseline model by 8.9%, due to over-segmentation. Compared to the Chichewa stemmer, we estimate that the unsupervised stemmer had 90-95% recall and 40-50% precision, resulting in over-stemmed tokens. However, this seemed to be favor the *segmented* and *morph-config* models, as unnecessary segmentation can be recovered when the tokens are sequenced or re-configured, with the supervised model arriving at the optimal weights for each candidate token or sequence. This can be seen by comparing the stemmed and morph-config results for the Chichewa-specific and language independent results. The difference in stemming is 20.7% but for the morph-config models it is only 2.8%. A loss in segmentation recall could not be recovered in the same way, as adjacent non-segmented morphemes will remain one token. This leads us to conclude that recall should be weighted more highly than precision in unsupervised morphological models applied to supervised classification tasks.

5.2 Phonology

For the phonological models the results in Table 2 show that the script-specific model was the most accurate with an average of F=0.443, a gain of 19.1% over the baseline.

There are correlations between morphological variation and phonological variation, with the gains similar for each label in Table 1 and Table 2. This is because much phonological variation often arises from the morphology, as in *ndiw̲odwala* where the glide *w* is pronounced and variably written between the vowels made adjacent through morphology. It is also because more morphologically complex words are longer and simply have more potential for phonological and written variation. The were greater gains in identifying the 'TB' and 'HIV' labels here than in the morphological models as the result of acronym identification.

The language independent model did not perform well. Despite changing the data considerably, there was little change in the accuracy, indicating that the changes it made were largely random with respect to the target concepts. The most frequent alternations in large contexts were noun-class prefixes differing by a single character, which has the potential to change the meaning, and this seemed to negate any gains from normalization.

While language independent results would have been ideal, a system with script-specific assumptions is realistic. It is likely that text messages are regularly sent in 1000s of languages but less than 10 scripts, and our definition of 'script specific' would be considered 'language independent' elsewhere. For example, in the Morpho Challenge (see

	Baseline	**Model**			**Gain**	
Label		Chichewa	Script	Indep	Best	Final
Patient-related	0.830	0.842	0.848	0.838	(+1.8)	(+1.8)
Clinic-admin	0.358	0.511	0.594	0.358	+23.6	+23.6
Technological	0	0.091	0.091	0	+9.1	+9.1
Response	0.113	0.420	0.473	0.207	+36.0	+36.0
Request for doctor	0.121	0.154	0.354	0	+23.3	+23.3
Medical advice	0	0.375	0.222	0.121	+37.5	+22.2
HIV	0.379	0.508	0.492	0.379	+12.9	+11.3
TB	0.235	0.327	0.492	0.235	+25.7	+25.7
Death	0.235	0.333	0.421	0.235	+18.6	+18.6
Average	0.252	0.396	0.443	0.264	+19.1	+19.1

Table 2: Phonological results: F-values for leave-one-out cross-validation comparing different phonological models. *Chichewa* = Chichewa specific heuristics, *Script* = specific to Roman scripts, *Indep* = language independent, ( ) = not significant ($\rho > 0.05, \chi^2$), *Final* = Gain of the 'Script' model over the Baseline.

Related Work) Arabic data was converted to Roman script, and it is likely that the methods could be adapted with some success to any alphabetic script.

5.3 Combined results

Table 3 gives the final results, comparing the systems over the original text messages and the English translations of the same messages. The most accurate results were achieved by applying the phonological normalization before the morphological segmentation, giving a (macro) average of 0.459 which is an increase of 20.6% over the baseline. The increase in accuracy was not cumulative – the combined system outperforms both the standalone phonological and morphological systems, but with a comparatively modest gain.

The final English system is 9.2% more accurate than the final Chichewa system, but the Chichewa system has closed the gap considerably as the English baseline system was 25.7% more accurate than the baseline Chichewa system. Assuming that the potential accuracy is approximately equal (given both languages are encoding exactly the same information) we conclude that we have made substantial gains in accuracy but there are further large gains to be made. Therefore, while we have not solved the problem of text message classification in morphologically rich languages, we have been able to make promising gains in an exciting new area of research.

5.4 Practical effectiveness

The FrontlineSMS system currently allows users to filter messages by keywords, similar to many email clients. Because of the large number of variants per word this is sub-optimal in many languages. We defined a second baseline to model an idealized version of the current system that assumes oracle knowledge of the keyword/label and the optimal order in which to apply rules created from this knowledge. The only constraint was that we excluded words that occurred only once. In essence, it is a MaxEnt model that includes seen test items and assigns a label according to the single strongest feature for each test item.

Here, we evaluated the systems according to Micro-F, recall and precision, as these give a better gauge of the frequency of error per incoming text, and therefore the usability for someone needing to correct mislabeled texts. We also calculated the Micro-F for each label/non-label decision to give exact figures per classification decision. The results are in Table 4. The Micro-F is 0.684 as compared to 0.403 for the keyword system. The higher precision is also promising, indicating that when we assign a label we are more often correct. By adjusting the precision and recall through label confidence thresholds, 90% precision can be achieved with 35.3% recall.[2] In terms of usability, the Label/no-Label re-

[2] We confirmed significance relative to confidence by ROC analysis – results omitted for space.

	Chichewa			English		
Label	**Baseline**	**Final Sys**	**Gain**	**Baseline**	**Final Sys**	**Gain**
Patient-related	0.830	**0.847**	(+1.7)	0.878	**0.878**	0
Clinic-admin	0.358	**0.624**	+26.6	0.682	**0.717**	(+3.4)
Technological	0	**0.174**	+17.4	0.174	**0.320**	+14.6
Response	0.113	**0.476**	+36.3	0.573	**0.555**	(-1.8)
Request for doctor	0	**0.160**	+16.0	0.160	**0.357**	+19.7
Medical advice	0.121	**0.500**	+37.9	0.560	**0.580**	(+2.0)
HIV	0.379	**0.357**	(-2.2)	0.414	**0.576**	+16.2
TB	0.235	**0.351**	+11.6	0.557	**0.533**	(-2.4)
Death	0.235	**0.638**	+40.3	0.591	**0.439**	-15.2
Average	0.252	**0.459**	+20.6	0.510	**0.551**	+4.1
Micro F	**0.593**	**0.684**	**+9.1**	**0.728**	**0.737**	**(+0.9)**

Table 3: Final Results, comparing the systems in Chichewa and the English translations.

sults are very promising, reducing errors from 1 in 4 to 1 in 20.

The learning rates in Figure 1 show that the learners are converging on accurate models after only seeing a handful of text messages. This figure also makes it clear that subword processing gives relatively little gain to the English translations. The disparity between the final model and the baseline widens as more items are seen, indicating that the failure of the word-optimal baseline model is not just due to a lack of training items.

5.5 Other models investigated

Much recent work in text classification has been in machine-learning, comparing models over constant features. We tested SVMs and joint learning strategies. The gains were significant but small and did not closed the gap between systems with and without subword modeling. We therefore omit these for space and scope.

However, one interesting result came from extending the feature space with topics derived from *Latent Dirichlet Allocation* (LDA) using similar methods to Ramage et al. (2009). This produced significant gains (micro-F=0.029), halving the remaining gap with the English system, but only when the topics were derived from modeling non-contiguous morpheme sequences, not words-alone or segmented morphemes. We found that the different surface forms of each word cooccurred *less* often than chance (0.46 as often as chance for the different forms of *odwala*) forming disjunctive distributions. We suspect that this acts as a bias against robust unsupervised clustering of the different forms.

6 Related Work

To our best knowledge, no prior researchers have worked on subword models for text message categorization, or any NLP task with the Chichewa, but we build on many recent developments in computational morphology and NLP for Bantu languages.

Badenhorst et al. (2009) found substantial variation in a speech recognition corpus for 9 Southern Bantu languages, where accurate models could also be built with limited data. Morphological segmentation improved Swahili-English machine translation in De Pauw et al. (2009), even in the absense of gold standard reference segmentations, as was the case here. The complexity and necessity of modeling non-contiguous morphemes in Bantu languages is discussed by Pretorius et al. (2009).

Computational morphology (Goldsmith, 2001; Creutz, 2006; Kurimo et al., 2008; Johnson and Goldwater, 2009; Goldwater et al., 2009) has begun to play a prominent role in machine translation and speech recognition for morphologically rich languages (Goldwater and McClosky, 2005; Tachbelie et al., 2009). In the current-state-of-the-art, a combination of the *ParaMor* (Monson et al., 2008) and *Morfessor* (Creutz, 2006) algorithms achieved

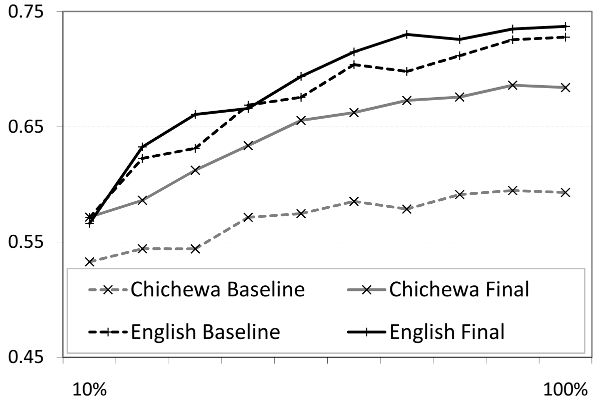

Figure 1: The learning rate, comparing micro-F for the Chichewa and English systems on different training set sizes. A random stratified sample was used for subsets.

the most accurate results in 2008 Morpho Challenge Workshop (Kurimo et al., 2008). *ParaMor* assumes a single affix and is not easily adapted to more complex morphologies, but we were able to test and evaluate *Morfessor* and the earlier *Linguistica* (Goldsmith, 2001). Both were more accurate for segmentation than our adaptation of Goldwater et al. (2009), but with lower recall. For the reasons discussed in Section 5.3 this meant less accuracy in classification. Goldwater et al. have also used the Pitman-Yor algorithm for morphological modeling (Goldwater et al., 2006). In results too recent to test here, Pitman-Yor has been used for segmentation with accuracy comparable to the HDP model but with greater efficiency (Mochihashi et al., 2009). Biosurveillance systems currently use simple rule-based pre-processing for subword models. Dara et al. (2008) found only modest gains, although the data was limited to English.

For text message classification, prior work is limited to identifying SPAM (Healy et al., 2005; Hidalgo et al., 2006; Cormack et al., 2007), where specialized algorithms and feature representations were also found to improve accuracy. For written variation, Kobus et al. (2008) focussed on SMS-specific abbreviations in French. Unlike their data, SMS-specific abbreviations were not present in our data. This is consistent with the reports on SMS practices in the related isiXhosa language (Deumert and Masinyana, 2008), but it may also be because the data we used contained professional communications not personal messages.

	Label class		**Label/No-Label**	
	KWF	Final	KWF	Final
F-val	0.403	0.684	0.713	0.950
Prec.	0.265	0.796	0.570	0.972
Rec.	0.842	0.599	0.953	0.929

Table 4: Micro-F, precision and recall, compared with the oracle keyword system. *KWF* = Oracle Keyword Filter.

7 Conclusions

We have demonstrated that subword modeling in Chichewa leads to significant gains in classifying text messages according to medical labels, reducing the error from 1 in 4 to 1 in 20 in a system that should generalize to other languages with similar morphological complexity.

The rapid expansion of cellphone technologies has meant that digital data is now being generated in 100s, if not 1000s, of languages that have not previously been the focus of language technologies. The results here therefore represent just one of a large number of potential new applications for short-message classification systems.

Acknowledgements

Thank you to FrontlineSMS:Medic and the health care workers they partner with. The first author was supported by a Stanford Graduate Fellowship.

References

Jaco Badenhorst, Charl van Heerden, Marelie Davel, and Etienne Barnard. 2009. Collecting and evaluating speech recognition corpora for nine Southern Bantu languages. In *The EACL Workshop on Language Technologies for African Languages*.

Piet Buys, Susmita Dasgupta, Timothy S. Thomas, and David Wheeler. 2009. Determinants of a digital divide in Sub-Saharan Africa: A spatial econometric analysis of cell phone coverage. *World Development*, 37(9).

Gordon V. Cormack, José Mara Gómez Hidalgo, and Enrique Puertas Sánz. 2007. Feature engineering for mobile (SMS) spam filtering. In *The 30th annual international ACM SIGIR conference on research and development in information retrieval*.

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

Jagan Dara, John N. Dowling, Debbie Travers, Gregory F. Cooper, and Wendy W. Chapman. 2008. Evaluation of preprocessing techniques for chief complaint classification. *Journal of Biomedical Informatics*, 41(4):613–23.

Ana Deumert and Sibabalwe Oscar Masinyana. 2008. Mobile language choices: the use of English and isiXhosa in text messages (SMS) evidence from a bilingual South African sample. *English World-Wide*, 29(2):117–147.

John Goldsmith. 2001. Unsupervised learning of the morphology of a natural language. *Computational Linguistics*, 27(2):153–198.

Sharon Goldwater and David McClosky. 2005. Improving statistical MT through morphological analysis. In *Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*.

Sharon Goldwater, Thomas L. Griffiths, and Mark Johnson. 2006. Interpolating between types and tokens by estimating power-law generators. *Advances in Neural Information Processing Systems*, 18.

Sharon Goldwater, Thomas L. Griffiths, and Mark Johnson. 2009. A bayesian framework for word segmentation: Exploring the effects of context. *Cognition*, 112(1):21–54.

Matt Healy, Sarah Jane Delany, and Anton Zamolotskikh. 2005. An assessment of case-based reasoning for Short Text Message Classification. In *The 16th Irish Conference on Artificial Intelligence & Cognitive Science*.

José Mara Gómez Hidalgo, Guillermo Cajigas Bringas, Enrique Puertas Sánz, and Francisco Carrero Garca. 2006. Content based SMS spam filtering. In *ACM symposium on Document engineering*.

Scott Isbrandt. 2009. Cell Phones in West Africa: improving literacy and agricultural market information systems in Niger. White paper: Projet Alphabétisation de Base par Cellulaire.

Abi Jagun, Richard Heeks, and Jason Whalley. 2008. The impact of mobile telephony on developing country micro-enterprise: A Nigerian case study. *Information Technologies and International Development*, 4.

Mark Johnson and Sharon Goldwater. 2009. Improving nonparameteric Bayesian inference: experiments on unsupervised word segmentation with adaptor grammars. In *Human Language Technologies*.

Catherine Kobus, François Yvon, and Geéraldine Damnati. 2008. Normalizing SMS: are two metaphors better than one? In *The 22nd International Conference on Computational Linguistics*.

Mikko Kurimo, Matti Varjokallio, and Ville Turunen. 2008. Unsupervised morpheme analysis. In *Morpho Challenge Workshop*, Finland. Helsinki University of Technology.

Carole Leach-Lemens. 2009. Using mobile phones in HIV care and prevention. *HIV and AIDS Treatment in Practice*, 137.

Sam Mchombo. 2004. *The Syntax of Chichewa*. Cambridge University Press, New York, NY.

Daichi Mochihashi, Takeshi Yamada, and Naonori Ueda. 2009. Bayesian unsupervised word segmentation with nested Pitman-Yor language modeling. In *The 47th Annual Meeting of the Association for Computational Linguistics*.

Christian Monson, Jaime Carbonell, Alon Lavie, and Lori Levin. 2008. ParaMor: finding paradigms across morphology. *Lecture Notes in Computer Science*, 5152.

Robert Munro. 2010. Haiti Emergency Response: the power of crowdsourcing and SMS. In *Haiti Crisis Relief 2.0*, Stanford, CA.

Steven Paas. 2005. *English Chichewa-Chinyanja Dictionary*. Mvunguti Books, Zomba, Malawi.

Guy De Pauw, Peter Waiganjo Wagacha, and Gilles-Maurice de Schryver. 2009. The SAWA Corpus: a parallel corpus of English - Swahili. In *The EACL Workshop on Language Technologies for African Languages*.

Gareth Peevers, Gary Douglas, and Mervyn A. Jack. 2008. A usability comparison of three alternative message formats for an SMS banking service. *International Journal of Human-Computer Studies*, 66.

Rigardt Pretorius, Ansu Berg, Laurette Pretorius, and Biffie Viljoen. 2009. Setswana tokenisation and computational verb morphology: Facing the challenge of a disjunctive orthography. In *The EACL Workshop on Language Technologies for African Languages*.

Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D. Manning. 2009. Labeled LDA: A supervised topic model for credit attribution in multilabeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, Singapore.

Ariel S. Schwartz and Marti A. Hearst. 2003. A simple algorithm for identifying abbreviation definitions in biomedical texts. In *The Pacific Symposium on Biocomputing*, University of California, Berkeley.

Martha Yifiru Tachbelie, Solomon Teferra Abate, and Wolfgang Menzel. 2009. Morpheme-based language modeling for amharic speech recognition. In *The 4th Language and Technology Conference*.

Yee Whye Teh, Michael I. Jordan, Matthew J. Beal, and David M. Blei. 2005. Hierarchical Dirichlet processes. *In Advances in Neural Information Processing Systems*, 17.