Typesetting for Improved Readability using Lexical and Syntactic Information

Ahmed Salama **Kemal Oflazer** **Susan Hagan**
Carnegie Mellon University – Qatar
Doha, Qatar
ahmedsaa@qatar.cmu.edu ko@cs.cmu.edu hagan@cmu.edu

Abstract

We present results from our study of which uses syntactically and semantically motivated information to group segments of sentences into *unbreakable units* for the purpose of typesetting those sentences in a region of a fixed width, using an otherwise standard dynamic programming line breaking algorithm, to minimize raggedness. In addition to a rule-based baseline segmenter, we use a very modest size text, manually annotated with positions of breaks, to train a maximum entropy classifier, relying on an extensive set of lexical and syntactic features, which can then predict whether or not to break after a certain word position in a sentence. We also use a simple genetic algorithm to search for a subset of the features optimizing F_1, to arrive at a set of features that delivers 89.2% Precision, 90.2% Recall (89.7% F_1) on a test set, improving the rule-based baseline by about 11 points and the classifier trained on all features by about 1 point in F_1.

1 Introduction and Motivation

Current best practice in typography focuses on several interrelated factors (Humar et al., 2008; Tinkel, 1996). These factors include typeface selection, the color of the type and its contrast with the background, the size of the type, the length of the lines of type in the body of the text, the media in which the type will live, the distance between each line of type, and the appearance of the justified or ragged right side edge of the paragraphs, which should maintain either the appearance of a straight line on both sides of the block of type (justified) or create a gentle wave on the ragged right side edge.

This paper addresses one aspect of current "best practice," concerning the alignment of text in a paragraph. While current practice values that gentle "wave," which puts the focus on the elegant look of the overall paragraph, it does so at the expense of meaning-making features. Meaning-making features enable typesetting to maintain the integrity of phrases within sentences, giving those interests equal consideration with the overall look of the paragraph. Figure 1 (a) shows a text fragment typeset without any regard to natural breaks while (b) shows an example of a typesetting that we would like to get, where many natural breaks are respected.

While current practice works well enough for native speakers, fluency problems for non-native speakers lead to uncertainty when the beginning and end of English phrases are interrupted by the need to move to the next line of the text before completing the phrase. This pause is a potential problem for readers because they try to interpret content words, relate them to their referents and anticipate the role of the next word, as they encounter them in the text (Just and Carpenter, 1980). While incorrect anticipation might not be problematic for native speakers, who can quickly re-adjust, non-native speakers may find inaccurate anticipation more troublesome. This problem could be more significant because English as a second language (ESL) readers are engaged not only in understanding a foreign language, but also in processing the "anticipated text" as they read a partial phrase, and move to the next line in the text, only to discover that they anticipated meaning incorrectly. Even native speakers with less skill may experience difficulty comprehending text and work with young readers suggests that "[c]omprehension difficulties may be localized at points of high processing demands whether from syntax or other sources" (Perfetti et al., 2005). As ESL readers process a partial phrase, and move to

the next line in the text, instances of incorrectly anticipated meaning would logically increase processing demands to a greater degree. Additionally, as readers make meaning, we assume that they don't parse their thoughts using the same phrasal divisions "needed to diagram a sentence." Our perspective not only relies on the immediacy assumption, but also develops as an outgrowth of other ways that we make meaning outside of the form or function rules of grammar. Specifically, Halliday and Hasan (1976) found that rules of grammar do not explain how cohesive principals engage readers in meaning making across sentences. In order to make meaning across sentences, readers must be able to refer anaphorically backward to the previous sentence, and cataphorically forward to the next sentence. Along similar lines, readers of a single sentence assume that transitive verbs will include a direct object, and will therefore speculate about what that object might be, and sometimes get it wrong.

Thus proper typesetting of a segment of text must explore ways to help readers avoid incorrect anticipation, while also considering those moments in the text where readers tend to pause in order to integrate the meaning of a phrase. Those decisions depend on the context. A phrasal break between a one-word subject and its verb tends to be more unattractive, because the reader does not have to make sense of relationships between the noun/subject and related adjectives before moving on to the verb. In this case, the reader will be more likely to anticipate the verb to come. However, a break between a subject preceded by multiple adjectives and its verb is likely to be more useful to a reader (if not ideal), because the relationships between the noun and its related adjectives are more likely to have thematic importance leading to longer gaze time on the relevant words in the subject phrase (Just and Carpenter, 1980).

We are not aware of any prior work for bringing computational linguistic techniques to bear on this problem. A relatively recent study (Levasseur et al., 2006) that accounted only for breaks at commas and ends of sentences, found that even those breaks improved reading fluency. While the participants in that study were younger (7 to 9+ years old), the study is relevant because the challenges those young participants face, are faced again when readers of any age encounter new and complicated texts that present words they do not know, and ideas they have never considered.

On the other hand, there is ample work on the basic algorithm to place a sequence of words in a typesetting area with a certain width, commonly known as the *optimal line breaking problem* (e.g., Plass (1981), Knuth and Plass (1981)). This problem is quite well-understood and basic variants are usually studied as an elementary example application of dynamic programming.

In this paper we explore the problem of learning where to break sentences in order to avoid the problems discussed above. Once such unbreakable segments are identified, a simple application of the dynamic programming algorithm for optimal line breaking, using unbreakable segments as "words", easily typesets the text to a given width area.

2 Text Breaks

The rationale for content breaks is linked to our interest in preventing inaccurate anticipation, which is based on the immediacy assumption. The immediacy assumption (Just and Carpenter, 1980) considers, among other things, the reader's interest in trying to relate content words to their referents as soon as possible. Prior context also encourages the reader to anticipate a particular role or case for the next word, such as agent or the manner in which something is done.Therefore, in defining our breaks, we consider not only the need to maintain the syntactic integrity of phrases, such as the prepositional phrase, but also the semantic integrity across syntactical divisions. For example, semantic integrity is important when transitive verbs anticipate direct objects. Strictly speaking, we define a bad break as one that will cause (i) unintended anaphoric collocation, (ii) unintended cataphoric collocation, or (iii) incorrect anticipation.

Using these broad constraints, we derived a set of about 30 rules that define acceptable and non-acceptable breaks, with exceptions based on context and other special cases. Some of the rules are very simple and are only related to the word position in the sentence:

- Break at the end of a sentence.
- Keep the first and last words of a sentence with the rest of it.

The rest of the rule set are more complex and depend on the structure of the sentence in question,

```
sanctions and UN charges of gross rights abuses. Military tensions on the
Korean peninsula have risen to their highest level for years, with the
communist state under the youthful Kim threatening nuclear war in response
to UN sanctions imposed after its third atomic test last month. It has also
```

(a) Text with standard typesetting

```
from US sanctions and UN charges of gross rights abuses. Military tensions
on the Korean peninsula have risen to their highest level for years,
with the communist state under the youthful Kim threatening nuclear war
in response to UN sanctions imposed after its third atomic test last month.
```

(b) Text with syntax-directed typesetting

Figure 1: Short fragment of text with standard typesetting (a) and with syntax and semantics motivated typesetting (b), both in a 75 character width.

e.g.:

- Keep a single word subject with the verb.
- Keep an appositive phrase with the noun it renames.
- Do not break inside a prepositional phrase.
- Keep marooned prepositions with the word they modify.
- Keep the verb, the object and the preposition together in a phrasal verb phrase.
- Keep a gerund clause with its adverbial complement.

There are exceptions to these rules in certain cases such as overly long phrases.

3 Experimental Setup

Our data set consists of a modest set of 150 sentences (3918 tokens) selected from four different documents and *manually* annotated by a human expert relying on the 30 or so rules. The annotation consists of marking after each token whether one is allowed to break at that position or not.[1]

We developed three systems for predicting breaks: a rule-based baseline system, a maximum-entropy classifier that learns to classify breaks using about 100 lexical, syntactic and collocational features, and a maximum entropy classifier that uses a subset of these features selected by a simple genetic algorithm in a hill-climbing fashion. We evaluated our classifiers *intrinsically* using the usual measures:

- *Precision*: Percentage of the breaks posited that were actually correct breaks in the gold-standard hand-annotated data. It is possible to get 100% precision by putting a single break at the end.
- *Recall*: Percentage of the actual breaks correctly posited. It is possible to get 100% recall by positing a break after each token.
- F_1: The geometric mean of precision and recall divided by their average.

It should be noted that when a text is typeset into an area of width of a certain number of characters, an erroneous break need not necessarily lead to an actual break in the final output, that is an error may not be too bad. On the other hand, a missed break while not hurting the readability of the text may actually lead to a long segment that may eventually worsen raggedness in the final typesetting.

Baseline Classifier We implemented a subset of the rules (those that rely only on lexical and part-of-speech information), as a baseline rule-based break classifier. The baseline classifier avoids breaks:

- after the first word in a sentence, quote or parentheses,
- before the last word in a sentence, quote or parentheses, and
- between a punctuation mark following a word or between two consecutive punctuation marks.

It posits breaks (i) before a word following a punctuation, and (ii) before prepositions, auxiliary verbs, coordinating conjunctions, subordinate conjunctions, relative pronouns, relative adverbs, conjunctive adverbs, and correlative conjunctions.

[1] We expect to make our annotated data available upon the publication of the paper.

Maximum Entropy Classifier We used the *CRF++ Tool*[2] but with the option to run it only as a maximum entropy classifier (Berger et al., 1996), to train a classifier. We used a large set of about 100 features grouped into the following categories:

- *Lexical features*: These features include the token and the POS tag for the previous, current and the next word. We also encode whether the word is part of a compound noun or a verb, or is an adjective that subcategorizes a specific preposition in WordNet, (e.g., *familiar with*).

- *Constituency structure features*: These are unlexicalized features that take into account in the parse tree, for a word and its previous and next words, the labels of the parent, the grandparent and their siblings, and number of siblings they have. We also consider the label of the closest common ancestor for a word and its next word.

- *Dependency structure features*: These are unlexicalized features that essentially capture the number of dependency relation links that cross-over a given word boundary. The motivation for these comes from the desire to limit the amount of information that would need to be carried over that boundary, assuming this would be captured by the number of dependency links over the break point.

- *Baseline feature*: This feature reflects whether the rule-based baseline break classifier posits a break at this point or not.

We use the following tools to process the sentences to extract some of these features:

- Stanford constituency and dependency parsers, (De Marneffe et al., 2006; Klein and Manning, 2002; Klein and Manning, 2003),

- lemmatization tool in NLTK (Bird, 2006),

- WordNet for compound nouns and verbs (Fellbaum, 1998).

[2]Available at `http://crfpp.googlecode.com/svn/trunk/doc/index.html`.

	Baseline	**ME-All**	**ME-GA**
Precision	77.9	87.3	89.2
Recall	80.4	90.2	90.2
$\mathbf{F}_1$	79.1	88.8	89.7

Table 1: Results from Baseline and Maximum Entropy break classifiers

Maximum Entropy Classifier with GA Feature Selection We used a genetic algorithm on a development data set, to select a subset of the features above. Basically, we start with a randomly selected set of features and through mutation and crossover try to obtain feature combinations that perform better over the development set in terms of F_1 score. After a few hundred generations of this kind of hill-climbing, we get a subset of features that perform the best.

4 Results

Our current evaluation is only intrinsic in that we measure our performance in getting the break and no-break points correctly in a test set. The results are shown in Table 1. The column ME-All shows the results for a maximum entropy classifier using all the features and the column ME-GA shows the results for a maximum entropy classifier using about 50 of the about 100 features available, as selected by the genetic algorithm.

Our best system delivers 89.2% precision and 90.2% recall (with 89.7% F_1), improving the rule-based baseline by about 11 points and the classifier trained on all features by about 1 point in F_1.

After processing our test set with the ME-GA classifier, we can feed the segments into a standard word-wrapping dynamic programming algorithm (along with a maximum width) and obtain a typeset version with minimum raggedness on the right margin. This algorithm is fast enough to use even dynamically when resizing a window if the text is displayed in a browser on a screen. Figure 1 (b) displays an example of a small fragment of text typeset using the output of our best break classifier. One can immediately note that this typesetting has more raggedness overall, but avoids the bad breaks in (a). We are currently in the process of designing a series of experiments for extrinsic evaluation to determine if such typeset text helps comprehension for secondary language learners.

4.1 Error Analysis

An analysis of the errors our best classifier makes (which may or may not be translated into an actual error in the final typesetting) shows that the majority of the errors basically can be categorized into the following groups:

- Incorrect breaks posited for multiword collocations (e.g., *act* of war*,[3] *rule* of law*, *far ahead* of*, *raining cats* and dogs*, etc.)
- Missed breaks after a verb (e.g., *calls* | *an act of war*, *proceeded to* | *implement*, etc.)
- Missed breaks before or after prepositions or adverbials (e.g., *the day after* | *the world realized*, *every kind* | *of interference*)

We expect to overcome such cases by increasing our training data size significantly by using our classifier to break new texts and then have a human annotator to manually correct the breaks.

5 Conclusions and Future Work

We have used syntactically motivated information to help in typesetting text to facilitate better understanding of English text especially by secondary language learners, by avoiding breaks which may cause unnecessary anticipation errors. We have cast this as a classification problem to indicate whether to break after a certain word or not, by taking into account a variety of features. Our best system maximum entropy framework uses about 50 such features, which were selected using a genetic algorithm and performs significantly better than a rule-based break classifier and better than a maximum entropy classifier that uses all available features.

We are currently working on extending this work in two main directions: We are designing a set of experiments to *extrinsically* test whether typesetting by our system improves reading ease and comprehension. We are also looking into a break labeling scheme that is not binary but based on a notion of "badness" – perhaps quantized into 3-4 grades, that would allow flexibility between preventing bad breaks and minimizing raggedness. For instance, breaking a noun-phrase right after an initial `the` may be considered very bad. On the other hand, although it is desirable to keep an object NP together with the preceding transitive verb, breaking before the object NP, could be OK, if not doing so causes an inordinate amount of raggedness. Then the final typesetting stage can optimize a combination of raggedness and the total "badness" of all the breaks it posits.

[3] * indicates a spurious incorrect break, | indicates a missed break.

Acknowledgements

This publication was made possible by grant NPRP-09-873-1-129 from the Qatar National Research Fund (a member of the Qatar Foundation). Susan Hagan acknowledges the generous support of the Qatar Foundation through Carnegie Mellon University's Seed Research program. The statements made herein are solely the responsibility of this author(s), and not necessarily those of the Qatar Foundation.

References

Adam Berger, Stephen Della Pietra, and Vincent Della Pietra. 1996. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):39–71.

Steven Bird. 2006. NLTK: The natural language toolkit. In *Proceedings of the COLING/ACL*, pages 69–72. Association for Computational Linguistics.

Marie-Catherine De Marneffe, Bill MacCartney, and Christopher D Manning. 2006. Generating typed dependency parses from phrase structure parses. In *Proceedings of LREC*, volume 6, pages 449–454.

Christiane Fellbaum. 1998. WordNet: An electronic lexical database. *The MIT Press*.

M. A. K. Halliday and R. Hasan. 1976. *Cohesion in English*. Longman, London.

I. Humar, M. Gradisar, and T. Turk. 2008. The impact of color combinations on the legibility of a web page text presented on crt displays. *International Journal of Industrial Ergonomics*, 38(11-12):885–899.

Marcel A. Just and Patricia A. Carpenter. 1980. A theory of reading: From eye fixations to comprehension. *Psychological Review*, 87:329–354.

Dan Klein and Christopher D. Manning. 2002. Fast exact inference with a factored model for natural language parsing. *Advances in Neural Information Processing Systems*, 15(2003):3–10.

Dan Klein and Christopher D. Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pages 423–430. Association for Computational Linguistics.

Donald E Knuth and Michael F. Plass. 1981. Breaking paragraphs into lines. *Software: Practice and Experience*, 11(11):1119–1184.

Valerie Marciarille Levasseur, Paul Macaruso, Laura Conway Palumbo, and Donald Shankweiler. 2006. Syntactically cued text facilitates oral reading fluency in developing readers. *Applied Psycholinguistics*, 27(3):423–445.

C. A. Perfetti, N. Landi, and J. Oakhill. 2005. The acquisition of reading comprehension skill. In M. J. Snowling and C. Hulme, editors, *The science of reading: A handbook*, pages 227–247. Blackwell, Oxford.

Michael Frederick Plass. 1981. *Optimal Pagination Techniques for Automatic Typesetting Systems*. Ph.D. thesis, Stanford University.

K. Tinkel. 1996. Taking it in: What makes type easier to read. *Adobe Magazine*, pages 40–50.