Computational Interpretations of Recency for the Choice of Referring Expressions in Discourse

Fahime Same
University of Cologne
f.same@uni-koeln.de

Kees van Deemter
Utrecht University
c.j.vandeemter@uu.nl

Abstract

First, we discuss the most common linguistic perspectives on the concept of recency and propose a taxonomy of recency metrics employed in Machine Learning studies for choosing the form of referring expressions in discourse context. We then report on a Multi-Layer Perceptron study and a Sequential Forward Search experiment, followed by Bayes Factor analysis of the outcomes. The results suggest that recency metrics counting paragraphs and sentences contribute to referential choice prediction more than other recency-related metrics. Based on the results of our analysis, we argue that, sensitivity to discourse structure is important for recency metrics used in determining referring expression forms.

1 Introduction

Speakers use various linguistic forms such as pronouns, proper names, and common nouns, to refer to entities in discourse. A great number of studies have addressed the issue of referring, and the factors that play a role in speakers' choice of the form of referring expressions. These factors include grammatical function (Brennan, 1995), animacy (Fukumura and van Gompel, 2011), competition (Arnold and Griffin, 2007), frequency (Ariel, 1990) and recency (McCoy and Strube, 1999; Ariel, 2001), among others. The focus of this article is on recency.

Broadly speaking, we understand recency to be the distance between the current mention of a referent and its antecedent. Therefore, in this work, we employ recency metrics to predict the form of subsequent mentions, and are not interested in the choice of "first-mention" expressions.

Recency has received much attention in both linguistic and computational studies, but in many cases, the notion of recency itself has been left largely undefined even though, as we shall see, recency can be understood in different ways. This paper has three objectives. The first is to survey different computational "interpretations" of the notion of recency. The second goal is to determine which of these computational interpretations is most effective for predicting the form of a referring expression in discourse context. In other words, we will ask, "what is the best way to operationalize the notion of recency in computational and data-oriented studies?" And the final objective is to see to which extent the choice of recency metrics should depend on the corpus.

The structure of this paper is as follows: in section 2, we summarize how recency has been used in linguistic studies. In section 3, we provide a brief overview of the notion of recency in Machine Learning (ML) studies, with the purpose of creating a taxonomy of recency metrics discussed in section 4. Sections 5 and 6 report two new studies. The former analyzes single recency metrics, the latter takes their combination into account. Finally, section 7 gives a brief summary and review of the findings.

2 Different interpretations of the notion of recency/distance

There is a long tradition of work in linguistics considering recency as a factor influencing the salience of a referent. The general idea is that the greater the distance between the two mentions, the greater the chance of using a full noun phrase anaphor (Vonk et al., 1992; Givón, 1992; Arnold, 2010); conversely, the shorter the distance between the two mentions, the greater the chance of pronominalization. Some studies have kept the notion of recency or "distance to the previous mention" opaque by not defining what long and short distance mean; while others have presented different interpretations of the notion of distance. In this paper, we focus on the three most frequent interpretations that are found in the literature.

Proceedings of the First Workshop on Computational Approaches to Discourse, pages 113–123

https://doi.org/10.18653/v1/P17

2.1 Immediate context

In the studies where the main focus is on the pronominalization problem, the notion of distance is often concerned with whether or not the antecedent is present in the same or previous utterance (or clause). In a corpus study, Hobbs (1978) noticed that in 98% of the cases, the antecedent of a pronoun anaphor is in the previous or in the same sentence. Ariel (1990) used the same sentence metrics in her corpus study, where she focused on the distribution of pronouns, demonstratives and full NPs. She demonstrated that with respect to distance from the antecedent, in more than 80% of cases, pronouns favor short distances, where the antecedent is in the same sentence or only one sentence away. In centering-based studies such as Hitzeman and Poesio (1998), Poesio et al. (2004) and Henschel et al. (2000) too, long distance antecedents are those which are more than one utterance or one clause away.

2.2 Non-local context

In some other corpus-based studies, a larger span of text was taken into account. In a comprehensive work on topic continuity in discourse, Givón (1983) measured the distance to the previous mention up to 20 clauses back. The work by Givón is one of the first attempts in quantifying the role of distance in discourse. In a computational pronominalization study, McCoy and Strube (1999) hypothesized that "when the last mention of an item is several sentences back in the text, a definite description is preferred". For this study which was conducted on a corpus of The New York Times articles, they found out that in long-distance situations (where the antecedent is more than two sentences away), a definite description is almost always used. In a psycholinguistics experiment, Arnold et al. (2009) examined the choice of referring expressions made by high-functioning children and adolescents with autism. Arnold et al. grouped the distance to the antecedent into 4 categories and demonstrated that the participants in their experiment had sensitivity to the discourse context.

2.3 Unit boundary

While the distance patterns explained in the previous paragraphs account for a large number of pronominalization cases, according to Fox (1987), they cannot handle all various types of anaphoric patterns. She showed that pronouns can be used to refer to a referent over long stretches of distance until the goal of the narrative changes (cited in Smith (2003)). In line with this idea, Ariel (1990) proposed the notion of *unity*, meaning, the antecedent being in the same frame, segment or paragraph. Vonk et al. (1992) and Tomlin (1987) also emphasized the importance of episode or unit boundaries, mostly realized as paragraph boundaries in written text, as factors contributing to the recency of mention.

As explained, there are three different interpretations of recency in the literature. The first two interpretations are concerned with measuring the distance in sentences (or clauses), while the third one goes beyond the sentential level, and focuses on paragraphs. Which of these interpretations does best in algorithms to predict referential choice in discourse contexts?

3 Recency in ML studies

Within Natural Language Generation (Gatt and Krahmer, 2018), reference production is computationally modelled in an area known as Referring Expression Generation (REG) (Krahmer and van Deemter, 2019; van Deemter, 2016). REG models have various shapes and forms, with feature-based ML models playing a substantial role.

GREC (Belz and Kow, 2010) was a series of Shared Task Evaluation tasks that is still regarded as a natural starting point when it comes to the generation of referring expressions in context. Different ML algorithms were submitted to these shared tasks, a number of which have exploited recency metrics. Some of the metrics used in these algorithms are pursuant to the interpretations mentioned in section 2. For example, the recency feature in Greenbacker and McCoy (2009) resembles the metric defined in McCoy and Strube (1999). Another example is a binary feature used by Bohnet (2008), which captures whether or not the antecedent occurs in the same sentence. This metric is similar to the interpretation discussed above under the heading "Immediate Context". Some of the other recency metrics used in these algorithms, however, are not in accordance with the interpretations introduced in section 2. For instance, Bohnet (2008) and Jamison and Mehay (2008) used distance metrics measuring number of words between the two mentions. In a more recent ML study, Kibrik et al. (2016) stated that referential choice belongs to a

large group of multifactorial processes. They used 7 different distance-related metrics in their study and concluded that these metrics are essential for successful prediction of referential choice, but there is no indication which metrics are the most relevant ones. Further studies that include recency metrics are Ferreira et al. (2016), Modi et al. (2017) and Saha et al. (2011), among others.

We saw that the metrics used in the ML studies are based on different units of measurement (e.g. word distance versus sentence distance). Likewise, different strategies are used to encode these metrics. For instance, some distances are measured in natural numbers while others are categorized in a smaller class of broader "bins". In the following example taken from the GREC-2.0 corpus (Belz et al., 2010), one could say that the distance between the expression "its" and its antecedent "Berlin" is 21 words (a natural number). Another solution would be, for instance, to follow Ferreira et al. (2016) in grouping the numerical distances into five groups consisting of 0-10 words, 11-20 words, 21-30 words, 31-40 words and more than 40 words. With this approach, the distance between "its" and its antecedent falls into the third bin, 21-30 words.

(1) **Berlin**$_{(1)}$ is$_{(2)}$ the$_{(3)}$ capital$_{(4)}$ city$_{(5)}$ and$_{(6)}$ one$_{(7)}$ of$_{(8)}$ the$_{(9)}$ sixteen$_{(10)}$ federal$_{(11)}$ states$_{(12)}$ of$_{(13)}$ Germany$_{(14)}$ ·$_{(15)}$ With$_{(16)}$ a$_{(17)}$ population$_{(18)}$ of$_{(19)}$ 3.4$_{(20)}$ million$_{(21)}$ in$_{(22)}$ **its**$_{(23)}$ city$_{(24)}$ limits$_{(25)}$,...

The question is which of these metrics work best in ML studies. The existing diversity motivated us to collect as many recency metrics as possible from the ML literature and create a taxonomy of recency metrics.

4 Methodology

This section begins with subsection 4.1 introducing recency metrics collected from different ML studies. Later, subsection 4.2 presents the two corpora used in our assessments and highlights their main differences. And finally, subsection 4.3 introduces the baseline algorithm and the ML method employed in our assessments.

4.1 Taxonomy of recency/distance metrics

Table 1 presents the metrics measuring the distance from the current expression to its antecedent [1]. As mentioned in the previous section, recency metrics vary a great deal. The most important differences between these metrics are:

I. Antecedent type In most metrics, the antecedent is the nearest previous mention of the same entity. In one of the metrics (metric 14 in Table 1), however, instead of the distance to the nearest mention, the distance to the nearest full NP mention is measured.

II. Unit of measurement The units in which the distance is measured vary in the recency metrics. The units of measurements used in the metrics outlined in Table 1 include distance in number of:

- words [metrics 1-3]
- sentences [metrics 4-11]
- NPs [metric 12]
- markables, defined as the textual expressions, between which coreferential relations can be established (Chiarcos and Krasavina, 2005). [metrics 13-14]
- paragraphs [metric 15]

III. Type of encoding As shown in Example (1), the major difference between encoding of the metrics is whether the distance is reported as a numeric value or defined bins. Among the metrics presented below, metrics 2, 3, 5, 6, 7 and 10 are categorical, the rest are numeric.

Another difference in type of encoding concerns how numeric values are encoded. Of the metrics used in this assessment, metrics 1, 4 and 12-15 are reported as natural numbers (including 0), metric 8 is the natural logarithm of the number of intervening sentences, metric 9 is its exponential variant [2] and metric 11, which will be explained below, is the normalized distance.

Scaled/normalized sentence distance The distance between the mentions ranges from 0 to 19 sentences in MSR and 0 to 146 sentences in WSJ. To overcome this sparsity, we decided to bound

[1] Greenbacker and McCoy defined the recency metric in their study as: "Referring expressions which were separated from the most recent reference by more than two sentences were marked as long distance references" (2009, p. 101). We have two different interpretations of this sentence which are presented as metric 5 and metric 6.

[2] The exponential distance is not reported for WSJ in this study.

Metric	Type of encoding & description	Meas Unit	Reference
1	Numerical distance	word	Bohnet (2008)
2	Categorical distance (5 bins of 0-10, 11-20, 21-30, 31-40 and 40+ words)	word	Ferreira et al. (2016)
3	Categorical distance (3 bins of 0-5, 6-12 and 13+ words)	word	Jamison and Mehay (2008)
4	Numerical distance	sentence	Orăsan and Dornescu (2009) Hendrickx et al. (2008) Kibrik et al. (2016) Saha et al. (2011)
5	Categorical distance [1st interp] (+/-2 sentences)	sentence	Greenbacker and McCoy (2009)
6	Categorical distance [2nd interp] (4 bins of 0,1,2,+ 2 sentences)	sentence	Greenbacker and McCoy (2009)
7	Categorical distance (3 bins of 0, 1, 2+ sentences)	sentence	Jamison (2008) Saha et al. (2011)
8	Log distance	sentence	Saha et al. (2011)
9	Exponential distance	sentence	Modi et al. (2017)
10	Antecedent in the same sentence?	sentence	Bohnet (2008)
11	Normalized distance	sentence	Newly implemented
12	Numerical distance	NP	Hendrickx et al. (2008)
13	Numerical distance	markable	Kibrik et al. (2016) Saha et al. (2011)
14	Numerical distance to the nearest non-pronominal antecedent	markable	Kibrik et al. (2016)
15	Numerical distance	paragraph	Kibrik et al. (2016)

Table 1: List of metrics collected from different ML studies

the values between two numbers [0,1], using the following formula:

$$x_{norm} = \frac{x_i - x_{min}}{x_{max} - x_{min}} \quad (1)$$

In this section, we introduced 14 metrics from the ML literature, plus one additional metric we decided to include in the study. The assessment of these metrics will be presented in section 5.

4.2 Corpora used in this study

As indicated earlier, we are also interested to find out the extent to which the choice of recency metrics should take the corpus itself into account. Corpora can be different from each other in terms of, for instance, size, genre (e.g. Wikipedia article, newspaper articles and medical reports) and structure of their documents (e.g. length and sentence structure). For this study, we have chosen two corpora which are different from each other in terms of *text genre* and length-related attributes (which will be referred to as *text structure* in this article).

Considering that the GREC Shared Tasks were among the first systematic studies tackling the referential choice in context, we decided to start our assessment of the metrics with GREC-2.0 (henceforth MSR[3]), one of the underlying corpora of these Shared Tasks[4]. MSR consists of more than 1500 introductory sections of Wikipedia articles in 5 different classes (people, city, country, river and mountain). The major pitfall of MSR is that only mentions to the main reference of the article are annotated.

In addition to MSR, we decided to include the Wall Street Journal portion (henceforth WSJ) of the OntoNotes corpus (Hovy et al., 2006; Pradhan et al., 2013) in this study. The genres of the two

[3] As this corpus is used in the GREC-MSR Shared Tasks, we abbreviate its name to MSR.

[4] We decided to exclude GREC-People, the other corpus used in these Shared Tasks because after the exclusion of the first mention expressions, only 121 instances of common nouns (2.16% of the whole data) were left. In a pilot study, we found out that the data is not enough for a three-way referential choice prediction task.

corpora are different, with the former containing Wikipedia articles, and the latter having newspaper articles. Also, the structure of the documents, such as length of each document, number of sentences and number of paragraphs are radically different across both corpora. The existing differences between the two corpora make it possible to explore whether the choice of recency metrics should depend on the text structure. Table 2 illustrates the major differences between the two corpora. In order to apply the recency metrics to MSR, we conducted tokenization and sentence segmentation using the spaCy python library. The texts of WSJ were already segmented and tokenized.

It is also important to note that four referring expression types, namely common noun, proper name, pronoun and zero anaphor are annotated in MSR. In WSJ, zero cases are not annotated, and only realized expressions are considered. For this reason, we decided to include only realized expressions (namely common nouns, proper names and pronouns) in our study and exclude the covert references. Hence, as mentioned before, the task in this study is to predict whether a target referring expression is a pronoun, a proper name or a common noun. The total number of referring expressions is 9306 in MSR and 21565 in WSJ, of which we placed 70% in a training set and 30% in a test set.

Corpus features	**MSR**	**WSJ**
number (n) of documents	1655	589
mean n of words / doc	166.5	600.8
mean n of sentences / doc	7.1	25
mean n of paragraphs / doc	2.3	10.8
mean n of chains / doc	n/a	15
mean length of sentences	25.8	29.5
n of common nouns	1613	6917
n of proper names	2813	7695
n of pronouns	4880	6953

Table 2: Comparison of the MSR and WSJ corpora in terms of length-related features and number of different types of referring expressions. Mean n of chains, meaning mean number of different annotated referents in a document, is not reported for MSR because only one chain per document is annotated.

As shown in Table 2, the documents in WSJ are roughly 4 times longer than the documents in MSR. Also, each document has a greater number of sentences and paragraphs. We expect that in the ML studies, the WSJ algorithms overall have a lower accuracy than the MSR algorithms.

4.3 Baseline algorithms and ML method

In order to assess the recency metrics, the first step is to create a baseline algorithm which contains no recency metric. This enables us to compare the performance of the experimental algorithms incorporating recency metrics against the baseline. We could have chosen different features, but we chose grammatical role of the current mention and grammatical role of the previous mention as the features of the baseline system for the following reasons: Using grammatical role is a safe choice, because the same syntactic categories were used in both corpora, so any differences in performance between the two corpora will not be due to differences in the annotations. Furthermore, we wanted to make sure that the features in the baseline algorithm are not confounding with recency metrics. For example, a competition-based feature such as the number of competing discourse entities between the two mentions would be confounding because the more competition there is, the greater the distance between the referent and the antecedent is likely to be. For this reason, we chose an algorithm that did not use anything other than grammatical role.

In this study, we use Multi-Layer Perceptron (henceforth MLP), a class of feedforward artificial neural networks as our ML approach. The model has two hidden layers with respectively 16 and 8 units. While hidden layers use the rectified linear activation function (ReLU), the output layer uses the softmax activation function. The model will be fit for 50 training epochs, and 50 samples (batch size) are being propagated through the network. It is noteworthy that since MLP cannot handle categorical data, all categorical metrics have been one-hot encoded in this study.

5 Assessing recency metrics using MLP

This section firstly reports on the success of the baseline algorithms, and continues with the algorithms incorporating the recency metrics.

5.1 Baseline algorithms

We mentioned in the previous section that the baseline algorithms are made up of two features, the grammatical role of the current mention and the grammatical role of its antecedent. Table 3 shows the accuracy of the two baseline algorithms.

	MSR	WSJ
baseline	0.585	0.55

Table 3: Accuracy of the MSR and WSJ baseline algorithms

5.2 Assessing recency metrics

Each experimental algorithm is composed of two baseline features and one recency metric. For instance, model 4 includes grammatical role of the current mention and the antecedent plus metric 4, which is the numerical distance in sentences. Since there are 15 different recency metrics and two different corpora, the total number of experimental algorithms is 30. If, for instance, an experimental algorithm would have 2 recency metrics instead of one, we would not be able to firmly test whether both features contribute to the performance of the algorithm, or only one of them is involved. For this reason, each metric is tested individually, and not in combination with other recency metrics. The overall accuracy of the experimental algorithms incorporating different recency metrics is reported in Table 4.

Meas Unit	Name	MSR	WSJ
Word	model 1	0.60	0.576
	model 2	0. 594	0. 551
	model 3	0.592	0. 572
Sentence	model 4	0.607	0.62
	model 5	0. 588	0. 582
	model 6	0.608	0. 622
	model 7	0.602	0.622
	model 8	0.607	0.611
	model 9	0.609	-
	model 10	0.589	0.597
	model 11	0.602	0.604
NP	model 12	0.59	0.623
Markable	model 13	-	0.577
	model 14	0.594	0.561
Paragraph	model 15	0.625	0. 616

Table 4: Accuracy of the experimental algorithms. The first column, Meas(urement) Unit specifies metrics' units of measurement detailed in section 4.1, II. Unit of measurement

The reported accuracies are all higher than the baseline accuracy, but it is still unclear whether the recency metrics are strongly informative of the probability of the increase in the accuracy of the algorithms.

We conducted Bayes Factor (henceforth BF) analysis using a beta distribution to investigate whether the outcomes of the experimental and the baseline algorithms come from distributions with the same underlying probability parameter, or ones with different underlying parameters. Hence, in the case of our current assessment, BF is used to determine whether or not there is good evidence for saying that the difference in accuracy rates of the models is less or greater than 0.01 (henceforth threshold). If the difference in accuracy is below the threshold, the evidence is in favor of similar distributions; if it is above the threshold, there is good evidence that the outcomes come from different distributions. In case of being from different distributions, we infer that the inclusion of recency metrics leads to an improvement in the performance of experimental algorithms.

Additionally, the strength of evidence for each experimental model versus the baseline will be assessed according to the scale of Kass and Raftery (1995).

BF	Interpretation
1 to 3	Not worth more than a bare mention
3 to 20	Positive
20 to 150	Strong
>150	Very strong

Table 5: Interpretation of Bayes Factors according to Kass and Raftery (1995, p. 777)

For the sake of space, we only report the results suggesting that the outcomes of the experimental and the baseline algorithms come from different distributions.

5.2.1 BF analysis of the MSR models

Comparing the rate of correct predictions of each experimental model to that of the baseline shows *positive* evidence that the accuracy of model 15, the one incorporating *distance in paragraph* as its recency metric, comes from different distribution than the baseline (BF=3.286). The other models were doing better than the baseline too, but there is insufficient evidence to say they are different from the baseline. More research is needed to investigate why other experimental models are not statistically different from the baseline.

5.2.2 BF analysis of the WSJ models

In the case of WSJ, the accuracy rates of 8 models are different from the accuracy of the baseline.

Similar to MSR, the outcome of model 15, utilizing the paragraph-based recency metric, comes from distributions with different underlying probabilities than the baseline. Additionally, *except* the outcome of model 5, there is very strong evidence that the accuracy of all other models (6 models in total) incorporating sentence-based recency metrics are being shifted by more than 0.01 beyond the baseline. This means, 6 out of 7 sentence-based recency metrics have improved the performance of the algorithms over the baseline. The remaining model with a different accuracy than the baseline is model 12, having NP distance as its recency metric.

Name	Meas	Def	BF
model 4	sentence	num	54×10^8
model 6	sentence	cat (4)	19×10^9
model 7	sentence	cat (3)	37×10^9
model 8	sentence	log	78×10^5
model 10	sentence	binary	14×10^2
model 11	sentence	norm	12×10^4
model 12	NP	num	56×10^9
model 15	paragraph	num	16×10^7

Table 6: Bayes Factor analysis giving the ratio of probabilities that the underlying accuracy rates are within 1% of each other or not. According to the scale of Kass and Raftery (1995) presented in Table 5, there is very strong evidence that the accuracy rates of all these models are different from the baseline. The column `Def` presents very briefly the definition of the metrics according to Table 1. For instance, `cat(4)` means the categorical distance in 4 bins.

5.2.3 BF analysis of the best performing models

As a next step, we compare the best performing models of each unit of measurement with each other. Since the only difference between the models is in their recency metrics, if there is good evidence that the difference in the accuracy of the models is greater than the threshold, we conclude that this difference is due to the differences in the recency metrics. Table 7 illustrates the best performing algorithms of each unit of measurement.

I. MSR models We conducted a one to one comparison between the best performing models of each unit. The evidence suggests that these models are not statistically different from each other.

II. WSJ models The evidence suggests that models 7, 12 and 15 are not evidentially distinguishable from each other. In other words, if we only focus on the WSJ corpus, we do not have enough evidence to prefer one model over another, and we can conclude that the best performing models incorporating sentence, paragraph and NP level recency metrics are equally good. But when we did a one to one comparison between these three models and the best performing models of word and markable units, we found out that the accuracy rates of each of these models have been shifted by more than 0.01 beyond the accuracy rates of the word and markable models. This means, the models incorporating paragraph, sentence and NP level metrics are statistically different from the models incorporating word and markable level information.

Meas Unit	MSR	WSJ
Word	Model 1	Model 1
Sentence	Model 9	Model 7
NP	Model 12	Model 12
Markable	Model 14	Model 13
Paragraph	Model 15	Model 15

Table 7: Best performing algorithms of each unit of measurement

As discussed in this section, the recency metrics clearly made a bigger improvement in the WSJ models. In the case of MSR, only one model had a distinguishable performance; while in the case of WSJ, 8 models performed statistically better than the baseline. Furthermore, sentence, paragraph and NP-based metrics evidentially improved the performance of the WSJ algorithms.

The results reported in this section were based on the assessment of single recency metrics; yet, there is no assessment of the combination of these metrics. In the next section, we report on a feature selection study we conducted to investigate which combinations of recency metrics lead to best results.

6 Sequential Forward Search

In order to investigate the extent to which the combination of different recency metrics improves the performance, we run a Sequential Forward Search (SFS) algorithm. The algorithm starts with an empty set and adds features to the model up to the point that no further improvement occurs. For this study, we used the R package mlr (Bischl et al., 2016) with the learner `classif.mlp`, and 5-fold cross-validation resampling strategy.

The result of the MSR experiment shows that the two recency metrics playing the most important roles are metric 15, distance in paragraph, and metric 9, exponential distance in sentences. Retraining the MLP algorithm on the new model, the accuracy is 0.637. The Bayes Factor analysis provides strong evidence that the outcome of this model is statistically different from the baseline (BF = 26.11).

In the WSJ SFS experiment, metric 15, distance in paragraphs, and metric 8, log distance in sentences, were chosen as the two recency features whose combination produced the best result. The model trained on the combination of these two metrics had the accuracy of 0.631. The Bayes Factor analysis finds very strong evidence that the outcomes of the baseline and this model are coming from different distributions.

What stands out in this experiment is that in the case of both MSR and WSJ, distance in paragraph is chosen as one of the recency metrics. The other chosen measures are exponential distance in MSR and logarithmic distance in WSJ. This could indicate that the algorithm is sensitive to the encoding of the sentence-based metrics. More experimentation in a more elaborated feature-based study is necessary to test this point.

7 Conclusion

Our goal was to shed light on different interpretations of recency, and to find out which of these interpretations are most effective for referential choice prediction. A subsidiary goal was to investigate whether the choice of recency metric should take corpus-specific features such as text genre and text structure into consideration.

The findings of this study should be of interest to theoretical and computational linguists alike, because both groups of researchers have studied the relation between recency and referential choice. In the linguistic tradition, the notion of recency has often been studied without a clear definition being offered (section 2). In the computational tradition, by contrast, researchers have dwelt less on theoretical justification but have had to provide precise definitions, to ensure that their algorithms are able to deal with a broad range of inputs. For example, Kibrik et al. (2016) defined 7 different implementations of the notion of recency taking different units of measurement into account; while Saha et al. (2011) employed various implementations of sentence-related metrics.

Another difference is that in the linguistic tradition, researchers usually think of recency as operating solely on the sentence or paragraph levels; while in computational works, less conventional metrics such as measuring the distance in words or NPs have been also practiced. We believe that the existence of a wider range of recency metrics in computational feature-based studies has the potential to open new windows into a better understanding of recency, and can encourage a re-evaluation of recency in the linguistic tradition. What is missing from many computational works is an explanation of why a certain metric or a certain way of encoding has been chosen over another. The findings from this study make the following contributions to the literature:

Creating a taxonomy of recency metrics After providing an overview of the most prevalent interpretations of recency in the linguistic tradition, we scrutinized the feature-based ML studies and provided, for the first time as far as we know, a taxonomy of recency metrics. The importance of this taxonomy is firstly that we do not know of any available work classifying and analyzing this notion comprehensively, so this work could be a starting point for getting deeper into the notion of recency.

Secondly, we have shed light on the differences between these metrics. Knowing what the differences are, and where they stem from, could be the first step in dissecting various aspects of this notion and developing new, improved recency metrics.

Assessing a wide range of recency metrics We have assessed individual metrics using the Multilayer Perceptron algorithm, and conducted a Bayes Factor analysis using a beta distribution to investigate whether there is evidence that the models incorporating recency metrics come from different distributions than the baseline algorithms. Additionally, we conducted a Bayes Factor analysis between the best performing models of each measurement unit to see whether there is enough evidence that the outcomes of models are different from each other.

The evidence reported in Table 6 for the models built on the WSJ corpus suggests that the outcome of the models incorporating NP, paragraph and sentence metrics have been shifted by more than 0.01 beyond the baseline's outcome. Also, we have strong evidence to believe that these models are sta-

tistically different from the models incorporating word and markable distance measures.

Additionally, the results of the Sequential Forward Search experiment show that, for both corpora, a combination of the paragraph-based and one of the sentence-based metrics leads to the best performance. This finding is important because it provides some direction in choosing recency metrics for feature-based computational studies. Furthermore, the Bayes Factor analysis and SFS combined suggest that "higher-level" metrics such as distance in paragraphs and sentences might result in greater changes in the performance of the algorithms than "lower-level" metrics based on counting words or markables. Finally, it raises the question of why a measurement such as distance in the number of sentences performs better than a measurement such as distance in the number of words. This is notable because the distance in words might be more indicative of the physical distance between the mentions, considering that sentences can vary enormously in length.

Another interesting observation is that some encoding solutions are more successful than others. For instance, the sentential distance in metric 5 is grouped into 2 bins of +/-2 sentences, while in metric 6, the distance is grouped into 4 bins of 0, 1, 2 or more than 2 sentences. While the former metric leads to a marginal difference in the performance of the algorithms, the latter contributes more to the improvement of the accuracy. These subtle differences in encoding and the great impact that they can make should be the focus of more experimentation.

Another major finding was the important role of distance measured in paragraphs. The Bayes Factor analysis showed that there is strong evidence for the differences between the performance of the baseline and the algorithms incorporating this metric. Also, using the SFS algorithm, this metric was selected in both MSR and WSJ as a feature contributing to the improvement of the results. The important role of paragraph information is in line with what we presented in section 2 under the topic "Unit boundary". According to Vonk et al. (1992), episode boundaries can decrease the accessibility of a referent, resulting in re-mentioning with full NPs. This might be the reason that including paragraph distance, and signaling whether or not the antecedent is in a different paragraph, makes the referential choice prediction simpler for the algorithms. The surprising point is that despite the major role of paragraph information, the only study from subsection 4.1 which has used the paragraph distance metric is Kibrik et al. (2016). The results from the current study could motivate a greater focus on paragraph-based information in feature-based studies.

Importance of the choice of corpus Surprisingly, the results of this study showed that recency measures were of greater importance when applied to WSJ than to MSR. In case of the MSR models, the only metric which in isolation led to a distribution different from the baseline was distance in the number of paragraphs, while in the case of WSJ, 8 different recency metrics led to major differences. One possible reason for the different behavior of recency metrics could be that due to unbalanced number of referring expression types (more than 50% pronouns and less than 20% common names), MSR is, most likely, not a suitable corpus for a three-way referential choice task.

It can be seen from the data in Table 2 that except the length of the sentences which is almost equal in both corpora, other text structure features, such as the number of words, sentences and paragraphs are very different from each other (with WSJ having almost 4 times more words, sentences and paragraphs). One speculation is that length-related features modulate the importance of the recency metrics in the ML models.

Further research is needed to identify the causes of this difference. However, based on our study, one might conclude that the more complex the discourse structure, the greater the role of recency measures. If this is true, it would be of great importance to carefully inspect the characteristics of the textual source prior to deciding which features to include in the study, as apparently, the choice of recency metric should depend on text genre and structure.

References

Mira Ariel. 1990. *Accessing Noun-Phrase Antecedents*. Routledge.

Mira Ariel. 2001. Accessibility theory: An overview. *Text representation: Linguistic and psycholinguistic aspects*, 8:29–87.

Jennifer E Arnold. 2010. How speakers refer: The role of accessibility. *Language and Linguistics Compass*, 4(4):187–203.

Jennifer E Arnold, Loisa Bennetto, and Joshua J Diehl. 2009. Reference production in young speakers with and without autism: Effects of discourse status and processing constraints. *Cognition*, 110(2):131–146.

Jennifer E Arnold and Zenzi M Griffin. 2007. The effect of additional characters on choice of referring expression: Everyone counts. *Journal of memory and language*, 56(4):521–536.

Anja Belz and Eric Kow. 2010. The GREC challenges 2010: overview and evaluation results. In *Proceedings of the 6th international natural language generation conference*, pages 219–229. Association for Computational Linguistics.

Anja Belz, Eric Kow, Jette Viethen, and Albert Gatt. 2010. Generating referring expressions in context: The task evaluation challenges. In *Empirical methods in natural language generation*, pages 294–327. Springer.

Bernd Bischl, Michel Lang, Lars Kotthoff, Julia Schiffner, Jakob Richter, Erich Studerus, Giuseppe Casalicchio, and Zachary M Jones. 2016. mlr: Machine Learning in R. *The Journal of Machine Learning Research*, 17(1):5938–5942.

Bernd Bohnet. 2008. IS-G: The comparison of different learning techniques for the selection of the main subject references. In *Proceedings of the Fifth International Natural Language Generation Conference*, pages 192–193. Association for Computational Linguistics.

Susan E Brennan. 1995. Centering attention in discourse. *Language and Cognitive processes*, 10(2):137–167.

Christian Chiarcos and Olga Krasavina. 2005. Annotation guidelines. pocos-potsdam coreference scheme. *Unpublished manuscript*.

Kees van Deemter. 2016. *Computational models of referring: a study in cognitive science*. MIT Press.

Thiago Castro Ferreira, Emiel Krahmer, and Sander Wubben. 2016. Towards more variation in text generation: Developing and evaluating variation models for choice of referential form. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 568–577.

Barbara A. Fox. 1987. *Discourse Structure and Anaphora: Written and Conversational English*. Cambridge Studies in Linguistics. Cambridge University Press.

Kumiko Fukumura and Roger PG van Gompel. 2011. The effect of animacy on the choice of referring expression. *Language and cognitive processes*, 26(10):1472–1504.

Albert Gatt and Emiel Krahmer. 2018. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research*, 61:65–170.

Talmy Givón. 1983. Topic continuity in discourse: An introduction. *Topic continuity in discourse: A quantitative cross-language study*, 3:1–42.

Talmy Givón. 1992. The grammar of referential coherence as mental processing instructions. *Linguistics*.

Charles Greenbacker and Kathleen McCoy. 2009. Udel: generating referring expressions guided by psycholinguistic findings. In *Proceedings of the 2009 Workshop on Language Generation and Summarisation*, pages 101–102. Association for Computational Linguistics.

Iris Hendrickx, Walter Daelemans, Kim Luyckx, Roser Morante, and Vincent Van Asch. 2008. Cnts: Memory-based learning of generating repeated references. In *Proceedings of the Fifth International Natural Language Generation Conference*, pages 194–195. Association for Computational Linguistics.

Renate Henschel, Hua Cheng, and Massimo Poesio. 2000. Pronominalization revisited. In *Proceedings of the 18th conference on Computational linguistics-Volume 1*, pages 306–312. Association for Computational Linguistics.

Janet Hitzeman and Massimo Poesio. 1998. Long distance pronominalisation and global focus. In *COLING 1998 Volume 1: The 17th International Conference on Computational Linguistics*.

Jerry R Hobbs. 1978. Resolving pronoun references. *Lingua*, 44(4):311–338.

Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. Ontonotes: the 90% solution. In *Proceedings of the human language technology conference of the NAACL, Companion Volume: Short Papers*, pages 57–60.

Emily Jamison. 2008. Using discourse features for referring expression generation. In *Proceedings of the 5th Meeting of the Midwest Computational Linguistics Colloquium (MCLC)*.

Emily Jamison and Dennis Mehay. 2008. Osu-2: Generating referring expressions with a maximum entropy classifier. In *Proceedings of the Fifth International Natural Language Generation Conference*, pages 196–197. Association for Computational Linguistics.

Robert E Kass and Adrian E Raftery. 1995. Bayes factors. *Journal of the american statistical association*, 90(430):773–795.

Andrej A Kibrik, Mariya V Khudyakova, Grigory B Dobrov, Anastasia Linnik, and Dmitrij A Zalmanov. 2016. Referential choice: Predictability and its limits. *Frontiers in psychology*, 7(1429).

Emiel Krahmer and Kees van Deemter. 2019. *Computational Generation of Referring Expressions: An Updated Survey*. Oxford University Press.

Kathleen F McCoy and Michael Strube. 1999. Generating anaphoric expressions: pronoun or definite description? In *The Relation of Discourse/Dialogue Structure and Reference*.

Ashutosh Modi, Ivan Titov, Vera Demberg, Asad Sayeed, and Manfred Pinkal. 2017. Modeling semantic expectation: Using script knowledge for referent prediction. *Transactions of the Association for Computational Linguistics*, 5:31–44.

Constantin Orăsan and Iustin Dornescu. 2009. WLV: A confidence-based machine learning method for the GREC-NEG'09 task. In *Proceedings of the 2009 Workshop on Language Generation and Summarisation (UCNLG+Sum 2009)*, pages 107–108. Association for Computational Linguistics.

Massimo Poesio, Rosemary Stevenson, Barbara Di Eugenio, and Janet Hitzeman. 2004. Centering: A parametric theory and its instantiations. *Computational linguistics*, 30(3):309–363.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Hwee Tou Ng, Anders Björkelund, Olga Uryupina, Yuchen Zhang, and Zhi Zhong. 2013. Towards robust linguistic analysis using ontonotes. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 143–152.

Sriparna Saha, Asif Ekbal, Olga Uryupina, and Massimo Poesio. 2011. Single and multi-objective optimization for feature selection in anaphora resolution. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 93–101.

Carlota S. Smith. 2003. *Referring expressions in discourse*, Cambridge Studies in Linguistics, page 123–152. Cambridge University Press.

Russell S Tomlin. 1987. *Coherence and grounding in discourse: outcome of a symposium, Eugene, Oregon, June 1984*, volume 11. John Benjamins Publishing.

Wietske Vonk, Lettica GMM Hustinx, and Wim HG Simons. 1992. The use of referential expressions in structuring discourse. *Language and cognitive processes*, 7(3-4):301–333.