Estimating Agreement by Chance for Sequence Annotation

Anonymous ACL submission

Abstract

The Kappa statistic is a popular chance corrected measure of agreement used as a reliability measure for annotation in the field of NLP, however its method for estimating chance agreement is not suitable for sequence annotation tasks, which are extremely prevalent in the field. The non-suitability is grounded in several complicating factors such as variation in span density across documents and constraints on span connectivity and overlap. In this paper, we propose a novel model for random annotation generation as the basis for chance agreement estimation for sequence annotation tasks. The model is jointly motivated by the specific characteristics of text sequence labeling tasks and acknowledgement of differences in annotation tendencies among annotators. Based on the proposed randomization model and related comparison approach, we successfully derive the analytical form of the distribution for computing the probable location of each annotated text segment, and subsequently chance agreement. We illustrate the approach in a simulation experiment and then apply it to several system outputs of CoNLL03 corpus annotation to evaluate its applicability, thus substantiating both the accuracy and efficacy of our method.

1 Introduction

Reliable annotation is an essential ingredient for NLP research agendas, both for enabling supervised learning methods, and also for evaluation. Though not frequently employed for evaluation of model performance in the field of NLP, one of the most widely accepted metrics for evaluation of annotation reliability is Cohen's Kappa, which offers an assessment of inter-rater reliability that is adjusted in order to avoid offering credit for the portion of observed agreement that can be attributed to chance. Some NLP tasks, such as Named Entity Recognition, and other span detection/labeling tasks, lack an appropriate chance corrected metric. This paper addresses that gap by proposing such a measure for these NLP tasks, illustrating its application in a simulation experiment, and then applying it to several system outputs of CoNLL03 corpus annotation.

Many previous studies have served as cautionary tails regarding the used of agreement measures that do not adjust for chance agreement, making the case that they might cause unfair comparisons among different tasks or systems since different tasks and systems are associated with different levels of chance agreement (Ide and Pustejovsky, 2017; Komagata, 2002; Gates and Ahn, 2017; Rand, 1971; Lavelli et al., 2008; Artstein and Poesio, 2008). Additionally, in the absence of a correction for the agreement by chance, the measurement values have a tendency to fall within a narrow range, which makes it more difficult to observe reliable differences between approaches (Eugenio and Glass, 2004). Therefore, estimating and correcting for chance agreement has become a critical step for annotation system evaluation, Apart from exceptional cases where chance agreement is small enough to be considered negligible.

From another angle, chance agreement is valuable apart from its role in estimating reliability in that it can also be used to quantify the difficulty of an annotation task. It is an important open problem to distinguish the difficulty of different annotation tasks, although we can qualitatively apply the intuition that large search spaces, large numbers of segments, and short segments usually correspond to more difficult sequence annotation tasks. Chance agreement allows us to quantitatively measure the difficulty level of different tasks and is consistent with human intuition.

The main contributions of our work are summarized as follows:

- We propose a novel random annotation model that incorporates different annotator tendencies, while taking into account the characteristics of processing each segment in the context of the

whole in sequence annotation task. The random annotation model can be further divided into two sub-models, which allows treating the case where overlap is allowed or not as separate cases. We also apply chance agreement to measure the difficulty of an annotation task. *To our knowledge, this is the first random annotation model that can be applied to complex text annotation tasks.*

- In view of the fact that many popular similarity measures are additive, the modeling of all dependent annotation segments in a text has been simplified to model each segment separately, although the location of each segment is still restricted by other segments in the same text. We successfully derive the analytical forms for the corresponding probability distributions of random annotations.

- We offer a simple formulation of the approach based on the discovery that the probabilities of most possible random locations for the same segment are the same, and thus we are able to offer a formalization that avoids redundant calculations. We discuss the asymptotic properties of the agreement by chance, and point out that chance agreement can be ignored when target annotations are sparse.

- We design and implement both simulation-based and naturalistic experiments. The experimental results demonstrate that our proposed method is accurate, effective and computationally efficient.

In the remainder of the paper we begin by laying our a theoretical framing for the work in a review of the past literature. We then explain our method. We first evaluate our method with a simulation study that enables broad exploration of the behavior of the approach and then follow up with applications to naturalistic corpora. We conclude with a discussion of limitations, ethical considerations, and future work.

2 Theoretical Foundation and Motivation

Estimation of chance agreement is a key element in the evaluation of classification tasks. However, though the field of NLP features a wide variety of span detection and labeling tasks, currently there is no widely adopted chance corrected metric for them.

In classification tasks, the Kappa coefficient is one of the most popular chance-corrected inter-annotator agreement measures (Komagata, 2002; Artstein and Poesio, 2008; Eugenio and Glass, 2004; Hripcsak and Rothschild, 2005; Powers, 2015; Cunningham and et al., 2014). The Kappa coefficient is defined as $(A_o - A_e)/(1 - A_e)$, where A_o is the observed agreement without chance agreement correction, and A_e is the expected agreement assuming random annotation. To estimate the chance agreement A_e, the key problem is how to build a random annotation model with reasonable assumptions. Many of the existing options assume independent annotations among different annotators, where each annotator is associated with some probability distribution that models the selection between categorical options (Artstein and Poesio, 2008).

Chance-corrected agreement is unarguably desirable for the evaluation of complex text annotation tasks beyond classification. Within this scope are structured prediction tasks, which include a plethora of information extraction tasks (Lampert et al., 2016; Esuli and Sebastiani, 2010; Dai, 2018). In these tasks, the problem of agreement by chance is more challenging to estimate than for simple classification tasks. In classification tasks, both the set of decisions that need to be made and the set of options for each of these decisions are consistent across annotators. For span prediction tasks, in contrast, annotators first identify spans that need to be labeled, and then choose a category for each span. Disagreement can occur at either level. It might be that they don't segment the text into the same spans, or they may assign the same span to a different label.

Take the Named Entity Recognition (NER) task (Cunningham and et al., 2014; Esuli and Sebastiani, 2010) as example, the number of entities and the length of each entity might vary widely among different annotators for the same text. Table 1 demonstrates a toy NER task annotated by two annotators. The text includes nine tokens that are each represented by one letter. The "Observed" column displays the observed annotations (highlighted text) for the same text from two annotators. Unlike simple annotation tasks, an annotator has the flexibility of choosing the number of entities and length of each entity. In this toy example, annotator 1 annotated two entities: one is "CDE" with 3 tokens, the other is "HI" with 2 tokens. Annotator 2 annotated only one entity of "EFGH" with 4 tokens.

	Observed	Random	Invalid Random
Annotator 1	ABCDEFGHI	ABCDEFGHI	ABCDEFGHI
Annotator 2	ABCDEFGHI	ABCDEFGHI	ABCDEFGHI

Table 1: Example of a Toy Named Entity Annotation.

Unfortunately, to our knowledge, there is currently no existing method for estimating agreement by chance for span prediction tasks like NER (Ide and Pustejovsky, 2017; Cunningham and et al., 2014). Although inter-annotator agreement estimation has become an important and necessary step for annotation evaluation, how to estimate chance agreement for complex text annotation is still a long-standing open problem. In line with this, and as pointed out by many previous studies, the sample space of a sequence annotation tasks like this is usually not well-defined (Ide and Pustejovsky, 2017; Cunningham and et al., 2014). For instance, we do not know how many non-entities exist in a text and do not even know how many tokens a non-entity should contain. Due to differences in annotation tendencies across annotators, the theoretical sample spaces are also different for individual annotators. Considering this variation in terms of differences in annotator preferences, some annotators like to merge adjacent information together, while others get used to labeling them as separate spans. Some annotators prefer to include ancillary surrounding text within a span, while others try to keep segments as short as possible. All of these factors make it challenging to estimate the agreement by chance for sequence annotation tasks.

While the specific problem of estimating chance agreement for span prediction tasks is an open problem, we must acknowledge that some relevant research has been done in connection with classification and clustering problems that informs our work and provides a continuum that our problem extends (Hennig et al., 2015; Fränti et al., 2014; Rezaei and Fränti, 2016; van der Hoef and Warrens, 2019; Warrens and van der Hoef, 2019; Meilă, 2007; Vinh et al., 2010). As mentioned, estimating agreement by chance is relatively simple in classification, because the sample space is fixed and the same for each annotator. In clustering problems, on the other hand, the situation is more challenging and somewhat more similar to that of span prediction problems. Conceptually, one might consider the elements that are within the same span might be analogous to elements within the same cluster. The most commonly used randomization model for clustering is the permutation model (Gates and Ahn, 2017), where all possible clusters with a fixed number of clusters and a fixed size of each cluster are randomly generated with equal probability. On the other hand, what makes span prediction different from the clustering case is that the permutation model in clustering does not put any constraints on the position of annotations in the same cluster. Annotations in the same cluster can be distributed anywhere. This assumption is not suitable in sequence annotations because annotations of the same segment are connected together and usually do not break into multiple fragments. In other words, annotators handle each segment as a whole, rather than labeling each token independently.

The variation in sample spaces caused by different labeling tendencies in light of connectivity constraints between segments make this problem quite challenging, especially when the annotated segments are required to be disjoint. Thus, in view of the characteristics of span prediction tasks and different annotation tendencies, we propose a new random annotation model that is compatible with these needs.

Our random annotation model first separately models the tendencies of each annotator. In particular, given observed annotations in a corpus, our random model learns how to perform the annotation task in a way that models tendencies without attending to the task-relevant characteristics that should distinguish between cases, thus uniformly randomizing the location of entities and preserving the annotator's distribution of segment lengths per category. Moreover, in order to meet the different requirements of various applications, we design two sub-models: the overlapping model and non-overlapping model in order to accommodate both the case where the task requires non-overlapping spans and the case where no such requirement is stipulated. For instance, the "Random" column in Table 1 shows an example of random annotation for each annotator. The random annotation for annotator 1 still has two entities: a 3-token one and a 2-token one with randomized locations. Invalid random examples are given at the "Invalid random" column in Table 1, since neither the number nor the length of entities are the same as the observed annotation. Note that the number of entities and

the length of each entity in the random annotation model are fixed for each annotator, but not being the same for all annotators for the same task. This is a common choice in the random annotation model because it reflects the different annotation tendencies of each annotator, which results in different chance agreements.

As a final motivating observation, we note that many similarity measures are additive. In other words, the comparison between the annotations of different annotators is an accumulation of comparisons between all pairs consisting of one labeled segment from one annotator with one from the other annotator. For example, the most popular metric F1 score for binary classification can be written as $2a/(2a+b+c)$, where a is the number of items labeled as positive by both annotators, b and c are the numbers of items rated as positive by one annotator but negative by the other (Hripcsak and Rothschild, 2005). Note that $2a+b+c$ is a constant when the number and length of spans are both observed. The rating of positive agreement a is the total number of positive agreements within the set of pairs of labeled segments with one from each annotator in a pair. We can simplify the modeling of the random sequence annotation by considering each segment separately instead of multiple ones together, even though each labeled segment is still subject to the constraints of other labeled segments in the same text if no overlap among segments is allowed. We successfully derive the analytical form for the distribution of the location of each single labeled segment. We also find that the probability is the same at most locations for each labeled segment, thereby avoiding a lot of redundant calculations. Details are presented in the next section.

3 Method

In this section, we first offer the specification of the random annotation model for sequence annotation, otherwise known as span prediction, then present the calculation, approximation, and asymptotic properties of chance agreement by random annotation. We focus on the most challenging non-overlapping models. Finally, we give a definition of the difficulty of an annotation task based on chance agreement. Due to space limitations, we only list the main conclusions and ideas in this section. For proof details, please refer to the appendix.

We adopt the named entity annotation (NER) as a representative of complex text sequence annotation tasks to demonstrate how to estimate the chance agreement or performance for sequence annotation evaluation. Given a text $T = \{t_1 \prec t_2 \prec \ldots \prec t_n\}$ with a sequence of n tokens $t_i, i \in \{1, \ldots, n\}$, and a pre-defined tag set $C = \{c_1, \ldots, c_m\}$ with m categorical tags; as a typical task in information extraction, named entity recognition aims to locate and classify segments of text T into pre-defined categories C, such as recognizing disease, medication, and symptom information from clinical notes. Mathematically, the annotation task for NER can be formulated as a function $\Phi : T \times C \mapsto \Omega$, where Ω is the set of all possible annotations. For any $\psi \in \Omega$, $\psi = \{\psi_{1,1}, \ldots, \psi_{1,k_1}, \ldots, \psi_{m,1}, \ldots, \psi_{m,k_m}\}$, where ψ is an annotation of segments for all pre-defined categories, k_i is the number of segments for i-th category. For an annotation segment $\psi_{i,j} = \{st_{i,j}, a_{i,j}\}$, $st_{i,j}$ denotes the index of the first token and $a_{i,j}$ denotes the length for the j-th segment with i-th category respectively. To simplify the discussion, in the following we will focus on single-tag text annotation (i.e., $m = 1$, $\psi = \{\psi_1, \ldots, \psi_k\}$, $\psi_j = \{st_j, a_j\}$) since it is straightforward to generalize the following techniques to multi-tag annotation as shown in the experiments.

To estimate the chance agreement, we need to define what we mean by random annotation. We cannot directly apply the permutation model for random clustering to sequence annotation tasks, because there is no constraint on the locations of annotations within the same cluster. The permutation model violates the intra-segment connectivity assumption that should hold in most text annotations. To tackle this problem, we propose a novel random annotation model that can take into account different annotators' choices and maintain the integrity of each text segment.

Random Sequence Annotation Model The random annotation model is a model that fixes the number and length of annotated segments for each annotator (but allows for differences across different annotators) and generates all possible annotation configurations with equal probability. In other words, for a k-segment random annotation $\Psi = \{\Psi_1, \ldots, \Psi_k\}$ with each randomly annotated segment $\Psi_i = \{ST_i, a_i\}$, it has equal probabilities for all possible start indices $\{st_1, \ldots, st_k\}$ with fixed lengths $a_1, \ldots, a_k$. For annotator 1 in Table

1, we have $k = 2$, $a_1 = 3$, $ST_1 \in \{1, \ldots, 7\}$, and $a_2 = 2$, $ST_2 \in \{1, \ldots, 8\}$. The definition of a random annotation segment $\{ST_i, a_i\}$ indicates its connectivity. All tokens in the same segment are consecutive without break and the index of the last token in the i-th annotated segment is $ST_i + a_i - 1$. In contrast, a random cluster generated by the permutation model for random clustering does not require this property. Note that the permutation of different entities is still allowed in our model as long as the segments within each entity remain contiguous, in other words, that the entity is permuted as a whole. As shown in the "Annotator 1" row of Table 1, different from the observed two entities with 3 and 2 tokens ("CDF" and "HI"), the left and right positions of the annotated entities in our random model with 3 and 2 tokens ("EFG" and "BC") can be swapped as illustrated in the "Random" column. With regards to different applications, the random annotation model can be further divided into two sub-models, namely, the overlapping model and the non-overlapping model. The overlapping model allows segments to overlap with each other, so each ST_i can take any value between 1 and $n - a_i + 1$, whereas the non-overlapping model does not allow segments to overlap, i.e., $ST_i \geq ST_j + a_j$ or $ST_j \geq ST_i + a_i$ for any $i \neq j$. Because the overlapping model is much easier to handle, and it can be easily derived from the non-overlapping one, we only focus on the non-overlapping model here.

The problem of estimating chance agreement for annotation evaluation can thus be described as follows:

Problem Definition. Assume there are two independent random annotations, $\Psi 1$ for annotator 1 and $\Psi 2$ for annotator 2 on the same text of length n. The problem is to estimate the expected similarity $E(Sim(\Psi 1, \Psi 2))$ based on a random non-overlapping annotation model.

Here we want to emphasize that for annotation of the same text, different annotators can label different numbers of text segments with different lengths. In this paper, we use right index instead of right subscript to represent the index of annotators, for example, $k1$ represents the number of segments annotated by annotator 1, and $k2$ for annotator 2. We notice that most agreement measures, regardless of being token level or entity level, can be formulated as segment-wise measures, i.e., $Sim(\psi 1, \psi 2) = f(\phi_{1,1}(\psi 1_1, \psi 2_1), \ldots, \phi_{k1,k2}(\psi 1_{k1}, \psi 2_{k2}))$, where $\psi 1_i = \{st1_i, a1_i\}$ is the i-th annotated segment for annotator 1 and $\psi 2_j = \{st2_j, a2_j\}$ is the j-th one for annotator 2. While it is challenging to estimate the chance agreement for a large number of dependent segments together with the random non-overlapping annotation model, the function f is additive for many popular measures. This fact allows us to process each segment individually, which greatly simplifies the estimation. We call the segment-wise measure with additive function f **additive measure**.

Proposition1 For the additive similarity measure, the expected chance agreement is $E(Sim(\Psi 1, \Psi 2)) = f(E\phi_{1,1}(\Psi 1_1, \Psi 2_1)), \ldots, E(\phi_{k1,k2}(\Psi 1_{k1}, \Psi 2_{k2})))$.

Note that in the non-overlapping random annotation model, the position of each random annotation segment is dependent on all the other random annotation segments within the same document from the same annotator. Since we assume all possible random annotations are equally likely, the problem of estimating the location distribution for each segment is equivalent to count the number of all possible configurations when we fix the location of the corresponding segment.

Proposition2 For the non-overlapping random annotation model, the number of all random annotations with the i-th segment fixed as:

$$\begin{aligned}&\Pi(ST_i = l) = \pi(l-1, 0)\pi(n-l-a+k, k-1)+\\&\sum_{i_1 \neq i} \pi(l - a_{i_1}, 1)\pi(n-l-a+a_{i_1}+k-1, k-2)+\\&\sum_{i_1 \neq i}\sum_{i_2 \neq i} \pi(l - a_{i_1} - a_{i_2} + 1, 2)\pi(n-l-a+a_{i_1}+a_{i_2}+k-2, k-3)\\&+ \ldots + \pi(l-a+a_i+k-2, k-1)\pi(n-l-a_i+1, 0),\end{aligned} \tag{1}$$

where $\pi(n, r) = n!/(n-r)!$ is the number of permutations of n things taken r at a time, k is the number of segments, a_i denotes the length of the i-th segment and $a = \sum_i a_i$ is the total length of annotations. Then the corresponding probability is $p(ST_i = l) = \Pi(ST_i = l)/\pi(n-a+k, k)$, for $1 \leq l \leq n - a_i + 1$. Here we treat each text segment as a different annotation, regardless of whether they have the same length. If we do not need to distinguish among entities of the same length, this formula can also be applied after a simple modification.

Proof sketch. We can divide all possible random annotations with $ST_i = l$ into k disjoint sets with m annotation segments located on the left of the specified i-th segment ψ_i and the remaining $k -$

$m-1$ segments on the right side. The cardinality of each set with selected left m annotation segments (which then determines the segments on the right) is the number of all possible annotations on the left $l-1$ times the number for $n-l-a_i$ of tokens on the right side.

If we fix the order of m selected random annotation segments ψ_{i_1}, ..., ψ_{i_m}, the random annotation of the left $l-1$ tokens is equivalent to distribute $l-1-\sum_{j=1}^{m} a_{i_j}$ objects into $m+1$ spaces, before the first annotation segment, between adjacent segments, and after the last one. This is a well studied problem (integer weak composition into a fixed number of parts) with $(l-1-\sum_{j=1}^{m} a_{i_j}+m)!/(l-1-\sum_{j=1}^{m} a_{i_j})!/m!$ possible configurations. Since we treat all annotation segments as different ones, there are $m!$ permutations for the left m segments and $(k-m-1)!$ for the right $k-m-1$ ones, and the cardinality of each set is $\pi(l-\sum_{j=1}^{m} a_{i_j}+m-1,m)\times\pi(n-l-a+\sum_{j=1}^{m} a_{i_j}+k-m,k-m-1)$. Based on the above derivation, the number of all possible configurations when we fix the location of a segment can be expressed by Equation 1.

However, it is computationally expensive to calculate Equation 1 for all possible random locations of each text segment when the sequence is very long. To solve this issue, we find that $\Pi(ST_i=l)$ is the same for most locations when the text is of length $n \gg a$.

Proposition3. ST_i is uniformly distributed for $a-a_i-k+2 \leq st_i \leq n-a+k$, i.e., $\Pi(st_i=l_1)=\Pi(st_i=l_2)$ for $\forall a-a_i-k+2 \leq l_1,l_2 \leq n-a+k$.

We further observe that it is not necessary to estimate chance agreement in all cases. Intuitively, we expect the chance agreement is small enough to be ignored when annotating sparse information in long texts and find that it is indeed the case. In most named entity recognition tasks, for example, the average tokens in annotated sentence is usually large than 20 (Roth and Yih, 2004).

Proposition4. The expected similarity $E(Sim(\Psi 1,\Psi 2)) \to 0$ when $n \gg a1+a2$, where $a1$ and $a2$ are the total lengths of all annotated segments for annotator 1 and annotator 2.

Proposition5. For the overlapping random annotation model, $p(ST_i=l)=1/(n-a_i+1)$, for $1 \leq l \leq n-a_i+1$.

In summary, we have proposed random sequence annotation models for both non-overlapping and overlapping cases. Under the condition of an additive similarity measure, we greatly simplify the estimation of the expected chance agreement in *Proposition* 1, and give its corresponding analytical formula in *Proposition* 2. We point out in *Proposition* 3 that each randomly annotated segment has the same probability for most locations except for a few ones at the left and right ends, which further reduces the computational cost. In addition, for long texts with sparse annotation information, the expected chance agreement is so small that it can be ignored, and this is proven in *Proposition* 4. The above conclusions are all for the case of the non-overlapping case, and the other case is very simple to deal with because it is subject to uniform distribution.

Another important application of chance agreement is to define the difficulty of an annotation task from the perspective of agreement by chance. Usually, evaluating the difficulty of annotation tasks is highly subjective and there are no good quantitative indicators. We utilize the chance agreement to define the difficulty of annotation tasks as follows:

Definition The difficulty level of an annotation task can be defined as $1-E(Sim(\Psi,\Psi))$ if there is a gold standard annotation Ψ or as average similarity of all annotator pairs $1-\sum_{i,j=1}^{v} E(Sim(\Psi 1,\Psi 2))/v^2$, where v is the number of annotators.

4 Experiments

In order to demonstrate the accuracy and efficacy of our method, we design simulation experiments for sequence annotations[1]. Since chance agreement estimation for the overlapping model is much simpler than chance agreement estimation for the non-overlapping model, all experiments in this paper are set up with the non-overlapping constraint.

Specifically, for the estimation of the probability distribution for random text annotation, we set to label four segments with lengths of 1, 5, 10, and 15 on a sequence of length 100. Figure 1 shows the probability distributions of the four segments at all possible locations calculated with the analytical formula in *Proposition* 2. The four distributions are approximately distributed as the inverted trapezoids with high ends and flat middle part, which confirms

[1]All experiments are implemented with MATLAB on a 2017 Mac Pro. The configuration of the Mac Pro is 2.9 GHz Intel Core i7 processor and 16GB 2133 MHz LPDDR3 memory. The evaluation tool and datasets will be released as open-source after the review period.

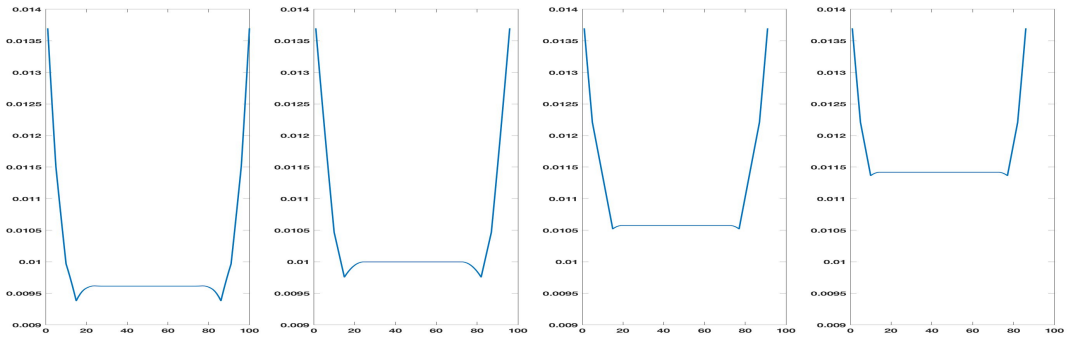

Figure 1: The probability distributions for all possible locations of each random segment in a length=100 sequence annotated with four segments. The lengths of the four segments are 1, 5, 10, 15, from left to right.

the conclusions of *Proposition* 2 and 3.[2]

The problem of chance estimation and correction is unique in that, to our knowledge, there is no real benchmark data that can be used to evaluate the performance. Therefore, most classic works in this field use synthetic data to illustrate and evaluate the effect of chance correction, such as Komagata (2002),Eugenio and Glass (2004), and Artstein and Poesio (2008). Intuitively, we know that the chance agreement is related to the size of the search space, the number of annotated objects, and the lengths of the annotated objects. We design the corresponding comparison experiments by varying these three factors.

We design three sets of comparison experiments by varying the length of text (simulation 1), the number (simulation 2) and length (simulation 3) of entities. In case *A* of simulation 1 shown in Table 2, we use 1 or 0 to indicate that each token in the text sequence is labeled or not. For the same sequence with 20 tokens, annotator 1 labels 3 entities with lengths of 2, 3, and 4. Annotator 2 labels 3 entities with lengths of 3, 4, and 5. The annotations of case *B* for two annotators are the same as in case *A*, the only difference is that ten 0s are added after the 20 tokens, that is, neither annotator 1 nor annotator 2 have labeled the extra 10 tokens. As reported in Table 4, because F1 score only focuses on the annotated tokens, the observed agreement (F1 score) is the same in both cases. However, since the labeled information in case *B* is relatively sparse, the chance agreement in case *B* is smaller, and the corresponding corrected F1 score is larger which means the agreement is higher. In simulation 2, the text length and the total number of annotated tokens remain the same, but the number of annotated entities changes from 3 in case A to 1 in case B. In simulation 3, the text length and the number of annotated entities remain the same, whereas the number of annotated tokens in case B is tripled. The results in Table 4, and 7 show that the longer the text, or the more entities, or the shorter the entities, the smaller the chance agreement and the higher the difficulty level. This is consistent with our intuition.

The main purpose of chance correction is to use different baselines for different tasks. In addition, chance correction may also change the ranking of model performance for the same task, although this is not common. As shown in the table 6, the gold standard annotation labels 6 entities with size of 3, 3, 3, 3, 3, 16. The annotator1 labels 5 3-token entities correctly but misses the 16-token entity. The annotator2 labels the 16-token entities correctly but misses 5 3-token entities. Note that the observed F1 score of annotator1 is lower than that of annotator2. But after the chance correction, the result is opposite (see table 7).

To evaluate our model in real data, we estimate the chance agreement of 11 state-of-the-art NER models (Liu et al., 2021) on CoNLL03 NER dataset (Sang and De Meulder, 2003), the results are shown in Table 8. Although it is difficult to validate chance agreement for real data without ground truth, we can see that the range of F1 scores is enlarged after the chance correction. We also divide the entire 3453 sentences of CoNLL03 data into two approximately equivalent subsets according to the difficulty level: subset1 with difficulty level less than 0.175 and subset2 with difficulty level greater than or equal to 0.175. As the results show, with respect to the performance ranking for 11 NER models, the rankings change significantly across different datasets. In addition, the performance ranking of all 11 models on subset2 also differs slightly before and after chance correction.

[2]The calculation time of the whole process is about 0.01 seconds.

	Observed (case A)	Observed (case B)
Annotator1	0 0 0 1 1 0 0 0 1 1 1 0 0 0 1 1 1 1 0 0	0 0 0 1 1 0 0 0 1 1 1 0 0 0 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0
Annotator2	0 0 1 1 1 0 0 0 1 1 1 1 0 0 1 1 1 1 1 0	0 0 1 1 1 0 0 0 1 1 1 1 0 0 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0

Table 2: Sequence Annotation Simulation 1.

	Observed (case A)	Observed (case B)
Annotator1	0 0 0 1 1 0 0 0 1 1 1 0 0 0 1 1 1 1 0 0	0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1 0 0 0 0 0
Annotator2	0 0 1 1 1 0 0 0 1 1 1 1 0 0 1 1 1 1 1 0	0 0 0 0 1 1 1 1 1 1 1 1 1 1 1 1 0 0 0 0

Table 3: Sequence Annotation Simulation 2.

Sim1	Obs F1	Chance F1	Cor F1	Diff Level	Sim2	Obs F1	Chance F1	Cor F1	Diff Level
CaseA	0.8571	0.5335	0.6938	0.4606	CaseA	0.8571	0.5335	0.6938	0.4606
CaseB	0.8571	0.3544	0.7787	0.6416	CaseB	0.8571	0.6455	0.5970	0.3478

Table 4: Chance Agreement Estimation for Sequence Annotation Simulation 1 and 2.

	Observed (case A)	Observed (case B)
Annotator1	0 0 0 0 0 0 0 0 1 1 1 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1 0 0 0 0 0
Annotator2	0 0 0 0 0 0 0 0 1 1 1 1 0 0 0 0 0 0 0 0	0 0 0 0 1 1 1 1 1 1 1 1 1 1 1 1 0 0 0 0

Table 5: Sequence Annotation Simulation 3.

Gold Standard	1 1 1 0 0 1 1 1 0 0 1 1 1 0 0 1 1 1 0 0 1 1 1 0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
Annotator1	1 1 1 0 0 1 1 1 0 0 1 1 1 0 0 1 1 1 0 0 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
Annotator2	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1

Table 6: Sequence Annotation Simulation 4.

Sim3	Obs F1	Chance F1	Cor F1	Diff Level	Sim4	Obs F1	Chance F1	Cor F1
CaseA	0.8571	0.1830	0.8251	0.8144	Anno1	0.6522	0.5013	0.3026
CaseB	0.8571	0.6455	0.5970	0.3478	Anno2	0.6808	0.5437	0.3005

Table 7: Chance Agreement Estimation for Sequence Annotation Simulation 3 and 4.

Model	F1-all				F1-subset1				F1-subset2				Time
	Obs	Rank	Cor	Rank	Obs	Rank	Cor	Rank	Obs	Rank	Cor	Rank	
A	0.923	3	0.901	3	0.919	2	0.911	2	0.9369	3	0.9035	4	23
B	0.905	7	0.878	7	0.889	7	0.878	7	0.9305	6	0.8938	6	23
C	0.9072	6	0.881	6	0.892	6	0.881	6	0.9320	5	0.8963	5	23
D	0.902	8	0.874	8	0.885	8	0.874	8	0.9261	7	0.8878	7	23
E	0.785	11	0.730	11	0.731	11	0.707	11	0.8537	11	0.7838	11	19
F	0.846	9	0.805	9	0.815	9	0.798	9	0.8929	9	0.8391	9	18
G	0.925	2	0.904	2	0.917	3	0.908	3	0.9414	2	0.9103	2	24
H	0.921	4	0.898	4	0.913	4	0.904	4	0.9368	4	0.9036	3	24
I	0.932	1	0.913	1	0.922	1	0.914	1	0.9500	1	0.9232	1	23
J	0.9073	5	0.882	5	0.903	5	0.894	5	0.9240	8	0.8851	8	22
K	0.802	10	0.752	10	0.759	10	0.737	10	0.8537	10	0.7854	10	16

Table 8: Chance Agreement Estimation for CoNLL03 Dataset. Obs is short for observed F1 as reported in corresponding real NER model (A-K), Cor is short for corrected F1. Time denotes the running time for chance estimation in seconds.

5 Conclusion and Discussion

In this paper, we propose a novel sequence random annotation model that takes into account the different annotation styles of annotators and the characteristics of sequence annotations. For complex cases where labeled objects are required to be disjoint, we investigate the corresponding distribution characteristic and remove redundant calculations. We also derive an analytical formula to calculate the exact distribution. Our focus in this work is how to establish a general framework and corresponding fast algorithm for calculating similarity by chance in complex text annotations. To the best of our knowledge, this is the first model and method that can calculate chance similarity for complex annotation tasks. The framework and method proposed in this paper are applicable to all additive similarity measures.

6 Limitations

Since this is the first work on chance estimation for sequence annotation, there is no other similar work to provide as a baseline for direct comparison. In addition, chance estimation lacks benchmark data with ground truth, although we have applied it to real data in order to demonstrate its utility. The current analysis of its effectiveness is mainly based on simulated data and whether it is consistent with human intuition. We expect that this work will stimulate more related work and benchmark data creation. The chance estimation in this paper focuses on the comparison between two annotators, and we plan to extend it to team-wise agreement for more than two annotators or systems.

7 Ethics Statement

The use of data on this project strictly adhered to ethical standards required by the National Institute of Health (NIH).

In addition to upholding ethical principles in conducting this work, we believe this work contributes to professional standards for rigor in the field. In particular, we expect that this paper will facilitate fair comparison of various annotation tasks or systems and reduce random chance agreement caused by different annotation styles and metrics. Chance agreement can also be used as a quantitative aid to measure the difficulty of annotation task. This provides a new perspective for evaluating different annotation tasks.

References

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Hamish Cunningham and et al. 2014. Developing language processing components with gate version 8.

Xiang Dai. 2018. Recognizing complex entity mentions: A review and future directions. In *Proceedings of ACL 2018, Student Research Workshop*, pages 37–44.

Andrea Esuli and Fabrizio Sebastiani. 2010. Evaluating information extraction. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 100–111. Springer.

Barbara Di Eugenio and Michael Glass. 2004. The kappa statistic: A second look. *Computational linguistics*, 30(1):95–101.

Pasi Fränti, Mohammad Rezaei, and Qinpei Zhao. 2014. Centroid index: cluster level similarity measure. *Pattern Recognition*, 47(9):3034–3045.

Alexander J Gates and Yong-Yeol Ahn. 2017. The impact of random models on clustering similarity. *The Journal of Machine Learning Research*, 18(1):3049–3076.

Christian Hennig, Marina Meila, Fionn Murtagh, and Roberto Rocci. 2015. *Handbook of cluster analysis*. CRC Press.

George Hripcsak and Adam S Rothschild. 2005. Agreement, the f-measure, and reliability in information retrieval. *Journal of the American medical informatics association*, 12(3):296–298.

Nancy Ide and James Pustejovsky. 2017. *Handbook of linguistic annotation*. Springer.

Nobo Komagata. 2002. Chance agreement and significance of the kappa statistic. *URL: http://www. tcnj. edu/komagata/pub/Kappa. pdf (Stand: Mai 2004)*.

Thomas A Lampert, André Stumpf, and Pierre Gançarski. 2016. An empirical study into annotator agreement, ground truth estimation, and algorithm evaluation. *IEEE Transactions on Image Processing*, 25(6):2557–2572.

Alberto Lavelli, Mary Elaine Califf, Fabio Ciravegna, Dayne Freitag, Claudio Giuliano, Nicholas Kushmerick, Lorenza Romano, and Neil Ireson. 2008. Evaluation of machine learning-based information extraction algorithms: criticisms and recommendations. *Language Resources and Evaluation*, 42(4):361–393.

Pengfei Liu, Jinlan Fu, Yang Xiao, Weizhe Yuan, Shuaicheng Chang, Junqi Dai, Yixin Liu, Zihuiwen Ye, Zi-Yi Dou, and Graham Neubig. 2021. Explainaboard: An explainable leaderboard for nlp. *arXiv preprint arXiv:2104.06387*.

Marina Meilă. 2007. Comparing clusterings—an information based distance. *Journal of multivariate analysis*, 98(5):873–895.

David MW Powers. 2015. What the f-measure doesn't measure: Features, flaws, fallacies and fixes. *arXiv preprint arXiv:1503.06410*.

William M Rand. 1971. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, 66(336):846–850.

Mohammad Rezaei and Pasi Fränti. 2016. Set matching measures for external cluster validity. *IEEE Transactions on Knowledge and Data Engineering*, 28(8):2173–2186.

Dan Roth and Wen-tau Yih. 2004. A linear programming formulation for global inference in natural language tasks. Technical report, ILLINOIS UNIV AT URBANA-CHAMPAIGN DEPT OF COMPUTER SCIENCE.

Erik F Sang and Fien De Meulder. 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition. *arXiv preprint cs/0306050*.

Hanneke van der Hoef and Matthijs J Warrens. 2019. Understanding information theoretic measures for comparing clusterings. *Behaviormetrika*, 46(2):353–370.

Nguyen Xuan Vinh, Julien Epps, and James Bailey. 2010. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *The Journal of Machine Learning Research*, 11:2837–2854.

Matthijs J Warrens and Hanneke van der Hoef. 2019. Understanding partition comparison indices based on counting object pairs. *arXiv preprint arXiv:1901.01777*.

8 Appendix

Proposition1 For the additive similarity measure, the expected chance agreement is $E(Sim(\Psi 1, \Psi 2)) = f(E\phi_{1,1}(\Psi 1_1, \Psi 2_1)), \ldots, E(\phi_{k1,k2}(\Psi 1_{k1}, \Psi 2_{k2})))$.

Proof.

Since the function f is additive, the order of the function f and expectation can be interchanged. We have $E(Sim(\Psi 1, \Psi 2)) = E(f(\phi_{1,1}(\Psi 1_1, \Psi 2_1), \ldots, \phi_{k1,k2}(\Psi 1_{k1}, \Psi 2_{k2}))) = f(E(\phi_{1,1}(\Psi 1_1, \Psi 2_1)), \ldots, E(\phi_{k1,k2}(\Psi 1_{k1}, \Psi 2_{k2})))$.

Originally, to estimate the expectation of similarity by chance, we need to sum up the similarity in a high-dimensional space of all possible random annotations, i.e., $E(Sim(\Psi 1, \Psi 2)) = \sum_{\Psi 1_1} \cdots \sum_{\Psi 1_{k1}} \sum_{\Psi 2_1} \cdots \sum_{\Psi 2_{k2}} f(.) \times p(\Psi 1_1 = \psi 1_1, \ldots, \Psi 2_{k2} = \psi 2_{k2})$. Now we can simplify it to multiple low-dimensional summations, such as $E(\phi_{i,j}(\Psi 1_i, \Psi 2_j))$, under the condition of additive measure.

Note that in the non-overlapping random annotation model, the position of each random annotation segment is dependent on all the other random annotation segments within the same document from the same annotator. Since we assume all possible random annotations are equally likely, the problem of estimating the location distribution for each segment is equivalent to count the number of all possible configurations when we fix the location of the corresponding segment.

Proposition3. ST_i is uniformly distributed for $a - a_i - k + 2 \leq st_i \leq n - a + k$, i.e., $\Pi(st_i = l_1) = \Pi(st_i = l_2) \, \forall \, a - a_i - k + 2 \leq l_1, l_2 \leq n - a + k$.

It is clear that proposition 3 and proposition 3* are equivalent.

Proposition3*. $\Pi(st_i = l) = \Pi(st_i = l+1) \forall a - a_i - k + 2 \leq l \leq n - a + k - 1$.

Proof sketch. Use mathematical induction

Initial step: when $k = 1$, $\Pi(st_1 = l) = 1$ and $p(st_1 = l) = 1/(n - a_1 + 1)$, for $1 \leq l \leq n - a_1 + 1$. So the proposition 3* is true at $k = 1$.

Inductive step: assume the proposition 3* holds for $k = r$. When $k = r + 1$, we partition all possible configurations with $st_i = l$ into $r + 1$ disjoint scenarios: the r scenarios with $st_j = l + a_i$ for all $j \neq i$ and the rest, i.e., the scenarios with a different annotation segment next to ψ_i from right side or none annotation segment next to ψ_i from right side. So $\Pi(st_i = l) = \sum_{j \neq i} \Pi(st_i = l \& st_j = l + a_i) +$

$\Pi(st_i = l \,\&\, st_j \neq l + a_i, \forall j \neq i)$.

We also partition all possible configurations with $st_i = l + 1$ into $r + 1$ disjoint scenarios: the r scenarios with $st_j = l + 1 - a_j$ for all $j \neq i$ and the rest, i.e., the scenarios with a different annotation segment next to ψ_i from left side or none annotation segment next to ψ_i from left side. Similarly, $\Pi(st_i = l + 1) = \sum_{j \neq i} \Pi(st_i = l + 1 \,\&\, st_j = l + 1 - a_j) + \Pi(st_i = l + 1 \,\&\, st_j \neq l + 1 - a_j, \forall j \neq i)$.

Since there is a bijection between the scenario of $st_i = l \,\&\, st_j \neq l + a_i, \forall j \neq i$ and the one of $st_i = l + 1 \,\&\, st_j \neq l + 1 - a_j, \forall j \neq i$ by identity mapping except the annotation segment ψ_i and the un-annotated token next to it with indices from l to $l + a_i$, $\Pi(st_i = l \,\&\, st_j \neq l + a_i, \forall j \neq i) = \Pi(st_i = l + 1 \& st_j \neq l + 1 - a_j, \forall j \neq i)$. For the pair of scenarios $st_i = l \,\&\, st_j = l + a_i$ and $st_i = l + 1 \,\&\, st_j = l + 1 - a_j$, they can be convert to scenarios $st_i^* = l \,\&\, a_i^* = a_i + a_j$ and $st_i^* = l + 1 - a_j \,\&\, a_i^* = a_i + a_j$ by merging ψ_i and ψ_j. Based on the assumption that the proposition 3* holds at $k = r$, their cardinalities should be equal since there is only r segments after the combination and $a - (a_i + a_j) - (k - 1) + 2 \leq l, l + 1 - a_j \leq n - a + (k - 1)$. Therefore, $\Pi(st_i = l \,\&\, st_j = l + a_i) = \Pi(st_i = l + 1 \,\&\, st_j = l + 1 - a_j)$ and the proposition 3* holds for $k = r + 1$.

It is a tight bound since we have to satisfy the condition of $0 \leq l - \sum_{j=1}^{m} a_{i_j} + m - 1$ and $0 \leq n - l - a + \sum_{j=1}^{m} a_{i_j} + k - m$ for all $0 \leq m \leq k - 1$ and $i_j \neq i$. This is the same as $a - a_i - k + 2 \leq l \leq n - a + k$.

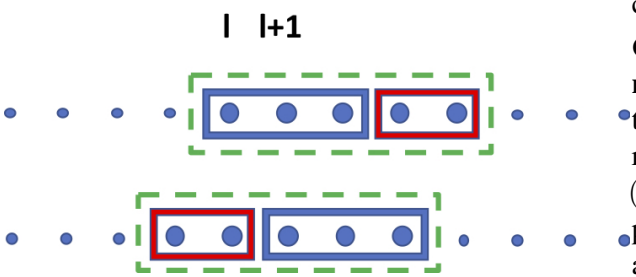

Figure 2: Convert the case of $k = r + 1$ to the case of $k = r$ by merging two adjacent text segments ψ_i and ψ_j, the blue box represents the segment ψ_i , and the red box represents the adjacent segment ψ_j.

Proposition4. The expected similarity $E(Sim(\Psi 1, \Psi 2)) \to 0$ when $n \gg a1 + a2$, where $a1$ and $a2$ are the total lengths of all annotated segments for annotator 1 and annotator 2.

Proof sketch. According to the proof process of *Proposition* 2, we know the number of all possible random annotations of k segments with total length a for a text with n tokens is $\pi(n - a + k, k)$. Thus, the total number of comparisons between random annotations from annotator 1 and annotator 2 is $\pi(n - a1 + k1, k1) \times \pi(n - a2 + k2, k2)$ under the independent annotation assumption. It is straight forward that the segment-wise agreement $\phi_{i_1,i_2}(\psi 1_{i_1}, \psi 2_{i_2})$ is zero if there is no overlap between the i_1-th text segment annotated by annotator 1 and the i_2-th text segment annotated by annotator 2. The agreement between two annotators is zero if there is no overlap among all $k1+k2$ annotated text segments. The situation is equivalent to combining the annotation results of the two annotators and requiring no overlap among all $k1+k2$ text segments in the same text. The total number of such possible annotations is $\pi(n - a1 - a2 + k1 + k2, k1 + k2)$. Therefore, the probability of zero chance agreement $p(Sim(\Psi 1, \Psi 2)) = 0) = \pi(n - a1 - a2 + k1 + k2, k1 + k2)/\pi(n - a1 + k1, k1)/\pi(n - a2 + k2, k2) = (n - a1 - a2 + k1 + k2) \times \ldots (n - a1 - a2 + 1)/((n - a1 + k1) \times \ldots (n - a1 + 1) \times (n - a2 + k2) \times \ldots (n - a2 + 1)) \to 1$ because both numerator and denominator are to the $(k1 + k2)$-th power of n and $n \gg a1 + a2 \geq k1 + k2$. Thus, we have $E(Sim(\Psi 1, \Psi 2)) \to 0$ when $n \gg a1 + a2$.

Proposition5. For the overlapping random annotation model, $p(ST_i = l) = 1/(n - a_i + 1)$, for $1 \leq l \leq n - a_i + 1$.

Proof sketch. This conclusion is straight forward because a random text segment annotation with length a_i can be placed at any feasible locations with equal probability without the non-overlapping constraint.

Computational complexity for random text annotation. The computational cost of calculating the probability distribution of the location of k random annotated text segments is bounded by $((k - 1) \times a - k^2 + 2k) \times 2^k \times (k - 1)$ multiplications and $((k - 1) \times a - k^2 + 2k) \times (2^k - 1)$ additions.

In order to calculate the probability distributions for random text annotation, according to the proposition 2 and the proposition 3, we could calculate the probability of $a - a_i - k + 2$ possible positions for each random annotated text segment with formula 1. And the analytical formula is a summation of 2^k terms, and each term is equivalent to $k - 1$ multiplications, so the computational complexity is bounded by $\sum_{i=1}^{k}(a - a_i - k + 2) \times 2^k \times (k - 1) = ((k - 1) \times a - k^2 + 2k) \times 2^k \times (k - 1)$ multipli-

cations and $\sum_{i=1}^{k}(a-a_i-k+2)\times(2^k-1) = ((k-1)\times a - k^2 + 2k)\times(2^k-1)$ additions. Since the formula 1 is a subset convolution, It may be possible to speed up this calculation with the fast subset convolution algorithm.

According to the above computational complexity analysis, we know that the probability distribution of the location of each random annotated segment can be calculated efficiently using the formula 1 when the number of text segments k is small. But with the increase of k, the computational cost will increase rapidly. Fortunately, when the text sequence is long enough and the annotated information is sparse, we can use the uniform distribution to approximate the distribution.

Uniform approximation. The probability distribution of the location of a random annotated text segment can be approximated by uniform distribution with $p(st_i = l) = 1/(n - a_i + 1)$, for $1 \leq l \leq n - a_i + 1$ if $(n-a+k)/(n-a_i+1) > \alpha$, where α is a preset threshold which is close to 1 and less than 1, for example $\alpha = 0.99$.

We observe that the probability distribution of the location of a random annotated text segment is approximately inverted trapezoid distributed with highest probabilities at both ends. And the majority of the whole distribution is flat when $n >> a$. It is straight forward to calculate the $p(st_i = 1) = \pi(n-a+k-1, k-1)/\pi(n-a+k, k) = 1/(n-a+k)$. So the distribution could be approximate with uniform distribution if the highest probability $1/(n-a+k)$ is close to the uniform probability $1/(n-a_i+1)$, i.e., $(n-a+k)/(n-a_i+1)$ is close to 1 if $n >> a$.

CoNLL03 NER dataset and system outputs. To evaluate our model in real data, we estimate the chance agreement of 11 state-of-the-art NER models on CoNLL03 NER dataset, the results are shown in Table 8. CoNLL-2003 is a named entity recognition dataset that is released as a part of CoNLL-2003 shared task: language-independent named entity recognition. This corpus consists of Reuters news stories between August 1996 and August 1997. There are four types of annotated entities: persons (PER), organizations (ORG), locations (LOC) and miscellaneous names (MISC). We downloaded 15 system outputs for the English test set from the Explained Board website after approval. Since 4 system outputs use different sentence segmentation, we limit our comparison to 11 system outputs that use the same sentence segmentation. The test set consists of 231 articles that include 3453 sentences.