Reconstructing Ancient Literary Texts from Noisy Manuscripts

Moshe Koppel [1] Moty Michaely [1] Alex Tal [2]

[1]Dept. of Computer Science, Bar Ilan University

[2]Dept. of Jewish Thought, University of Haifa

January 30, 2024

- Before printing techniques, all writing was done by hand.
- **Scribes** copied these works to the best of their abilities.
  $\rightarrow$ This naturally led to many **imperfect copies**.

Textual criticism

Explanation of Textual criticism by Maas [1958].

> *We have no autograph [handwritten by the original author] manuscripts of the Greek and Roman classical writers and no copies which have been collated with the originals; the manuscripts we possess derive from the originals through an unknown number of intermediate copies, and are consequently of questionable trustworthiness. The business of textual criticism is to produce a text as close as possible to the original (constitutio textus).*

Example from the Hebrew bible[1]

Genesis 1:2,

- 'and the earth was formless and void,'
- 'but the earth was unseen and unready'
- 'But the earth was lifeless and empty'

[1] Source: https://en.wikipedia.org/wiki/Textual_variants_in_the_Hebrew_Bible

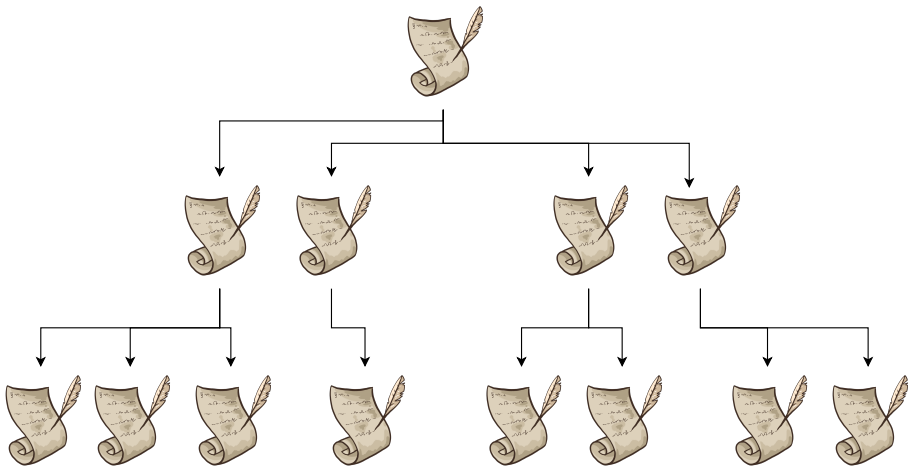
- Of course historians try to preserve the original texts.
- How can we get as close to those original texts as possible?

Manual reconstruction approaches

- Select the copy that best represents the original text. - Stemmatic approach
- Collate a new text from the various copies that best represents the original text.

Stemmatic approach

- For the stemmatic approach the challenge is to create the stemma, a tree diagram that shows which text was transcribed from which. The root of the tree shows the original text.



- The "Stemmatic" approach is preferable when the collection of extant manuscripts for a given text is relatively complete
- Especially if the original text is found in the collection.
- In the case of ancient documents, this situation is very rare.

- For many historical manuscripts, the problem is that the original text is not in our collection and the collection is heavily limited in size.
- A new idea is needed to alleviate this problem.

1. Align all the texts to each other.
2. Cluster related texts together.
3. Using statistical methods, judge which words from which aligned text to take to be as close to the ur-text as possible.

Aligning the texts - synopsis

- To reconstruct the original text, we first need to arrange all manuscripts so that parallel words or phrases can be compared.

Aligning the texts

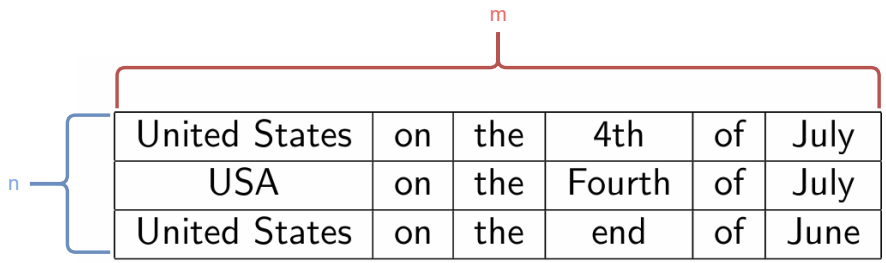
United States	on	the	4th	of	July
USA	on	the	Fourth	of	July
United States	on	the	end	of	June

Aligning the texts

- This process can be done by hand or automatically.
- In their research, the authors profit from manually created synopses.

Formulating the problem

- We have a n x m synopsis matrix $a = \{a_{ij}\}$ where a_{ij} is one cell of the matrix. n is the number of manuscripts and m the number of words/phrases.



Formulating the problem

- For each column a_j, there is one correct token and k_j distinct tokens other than the correct token, so in total $k_j + 1$.
- We map each choice a_{ij} to a number in the distinct tokens set $\{1, ...k_j + 1\}$. We denote t_j the number of the correct token.
- In our example below: We map *United States* to 1 and consider it as the correct token t_1, and we have one other distinct form *USA* which we map to 2, so $k_1 = 1$.

United States→ 1	on→ 1	the→ 1	4th→ 1	of→ 1	July→ 1
USA→ 2	on→ 1	the→ 1	Fourth→ 2	of→ 1	July→ 1
United States→ 1	on→ 1	the→ 1	end→ 3	of→ 1	June→ 2

Formulating the problem

- Each document (= row a_i) has a reliability probability p_i. It denotes the probability that the scribe correctly transcribes a manuscript.
- We have a document reliability set $\{p_i\}_i$ containing all the document reliability probabilities.

p_1	United States	on	the	4th	of	July
p_2	USA	on	the	Fourth	of	July
p_3	United States	on	the	end	of	June

Formulating the problem

- We consider an urtext reconstruction attempt a mapping from the synopsis matrix $a = \{a_{ij}\}$ to a proposed text in our sets of distinct forms $\{1, ...k_j + 1\}^m$
- The goal is to find an optimal reconstruction given only the synopsis matrix a

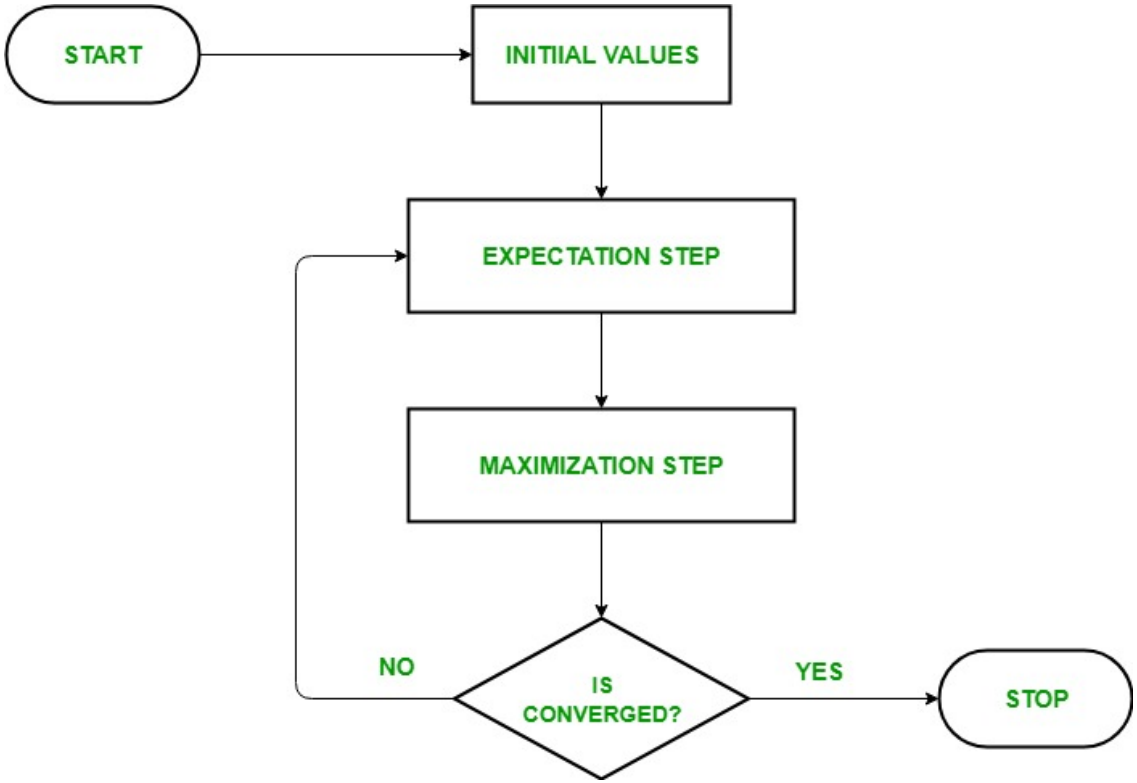
United States$\rightarrow 1$	on$\rightarrow 1$	the$\rightarrow 1$	4th$\rightarrow 1$	of$\rightarrow 1$	July$\rightarrow 1$
USA$\rightarrow 2$	on$\rightarrow 1$	the$\rightarrow 1$	Fourth$\rightarrow 2$	of$\rightarrow 1$	July$\rightarrow 1$
United States$\rightarrow 1$	on$\rightarrow 1$	the$\rightarrow 1$	end$\rightarrow 3$	of$\rightarrow 1$	June$\rightarrow 2$

Formulating the problem

- Optimality is obtained by values $\{p_i\}_i$ and $p(t_j = w | w \in \{1, \ldots, k+1\})$ (or for short $\{p(t_j = w)\}_j$) that maximize the likelihood of a.
- $\{p(t_j = w)\}_j$ can be computed using a and $\{p_i\}_i$. Thus, we must maximize $p(a; \{p_i\})$.

Expectation Maximization algorithm

The authors use a modified expectation maximization algorithm.



UR Algorithm

- We assign an initial constant value to $\{p_i\}_i$, then follow these two steps until convergence:
  1. We use the p_i values to update the probabilities $\{p(t_j = w)\}_j$.
  2. We update the p_i values using $\{p(t_j = w)\}_j$.

UR Algorithm - First step

We update $\{p(t_j = w|a)\}$ for each column a_j and for each $w \in \{1, \ldots, k_j + 1\}$

$$\{p(t_j = w|a)\} = \{p(t_j = w|a_j)\} = \frac{\{p(a_j|t_j = w)\}}{Z} \tag{1}$$

$$= \frac{\prod_{a_{ij}=w} p_i \cdot \prod_{a_{ij}\neq w}(1 - p_i)/k_j}{Z} \tag{2}$$

Example

1. Set p_i to some values
2. Perform expectation step, for w = "United States":
   $p_1 \cdot p_3 \cdot (1 - p_2)/k_1$

p_1	United States	on	the	4th	of	July
p_2	USA	on	the	Fourth	of	July
p_3	United States	on	the	end	of	June

UR Algorithm - Second step

- We compute the maximum-likelihood values of $\{p_i\}_i$ by comparing $\{p(t_j = w|a)\}$to the judgements of individual i.
- The intuition is that the maximum likelihood value of p_i is equal to the average probability that $a_{ij} = t_j$ The new updated value of p_i is therefore:

$$p_i = \frac{1}{m}(\sum_j p(t_j = a_{ij}|a)) \tag{3}$$

- UR algorithm is optimal only when manuscripts are independent of each other.
- However, manuscripts were copied from one another and thus can't be independent.

Handling lack of independency

- The idea is to cluster manuscripts that show similar errors, then use the UR algorithm to identify the original text for each cluster.
- The authors do not look for automatic clustering methods, as they state that domain experts should be able to cluster these texts.
- They denote this method recursive UR.

- The authors test the UR algorithm on 3 different groups of manuscripts:
  1. Artificial manuscripts – 2nd generation copies.
  2. Artificial manuscripts – 3rd generation copies.
  3. Two Real-World examples.
- Simple Majority Rule is used as the baseline.

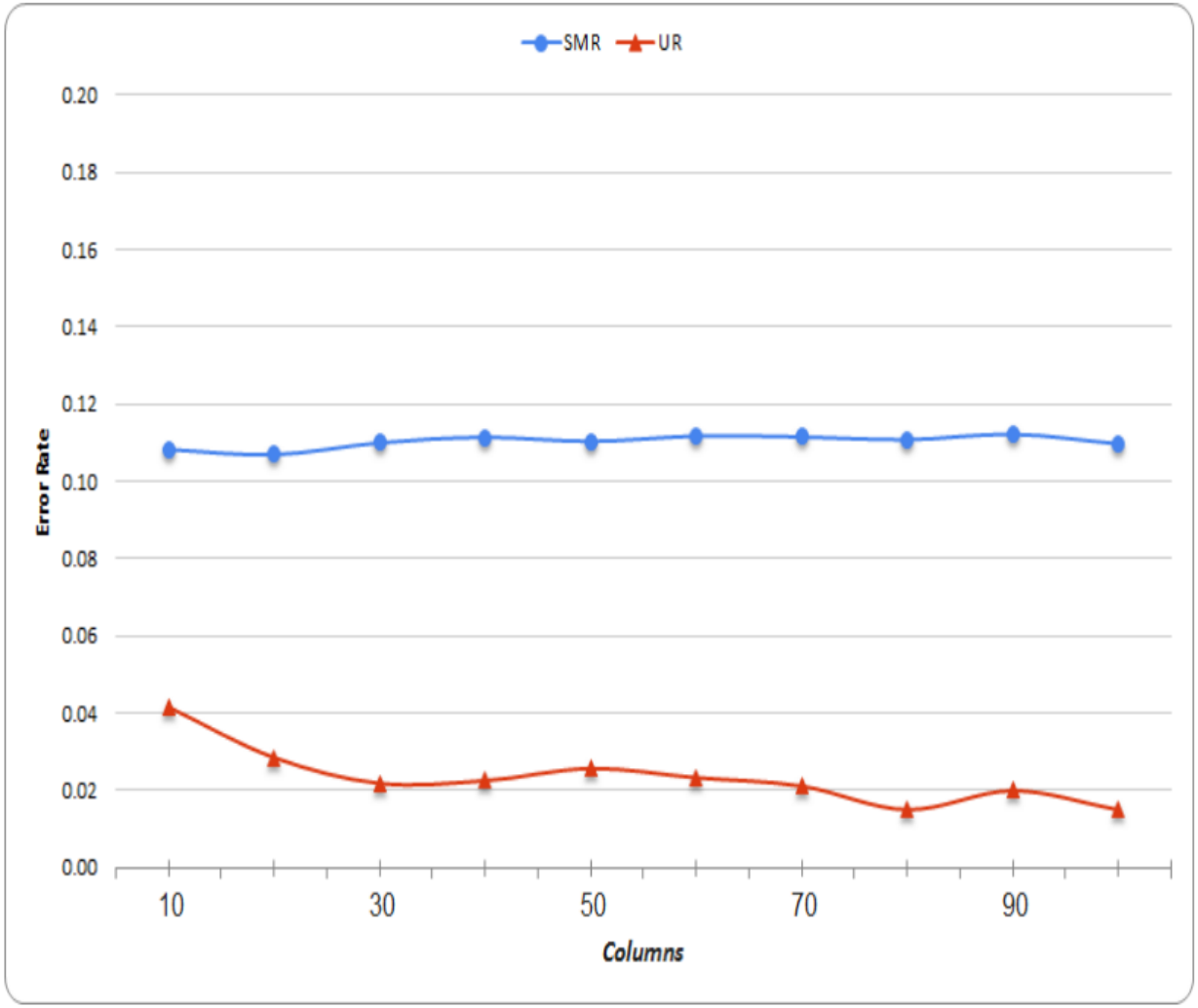
Baseline

- Simple Majority Rule (SMR)
- SMR chooses the distinct form in a column of the synopsis with the highest count.

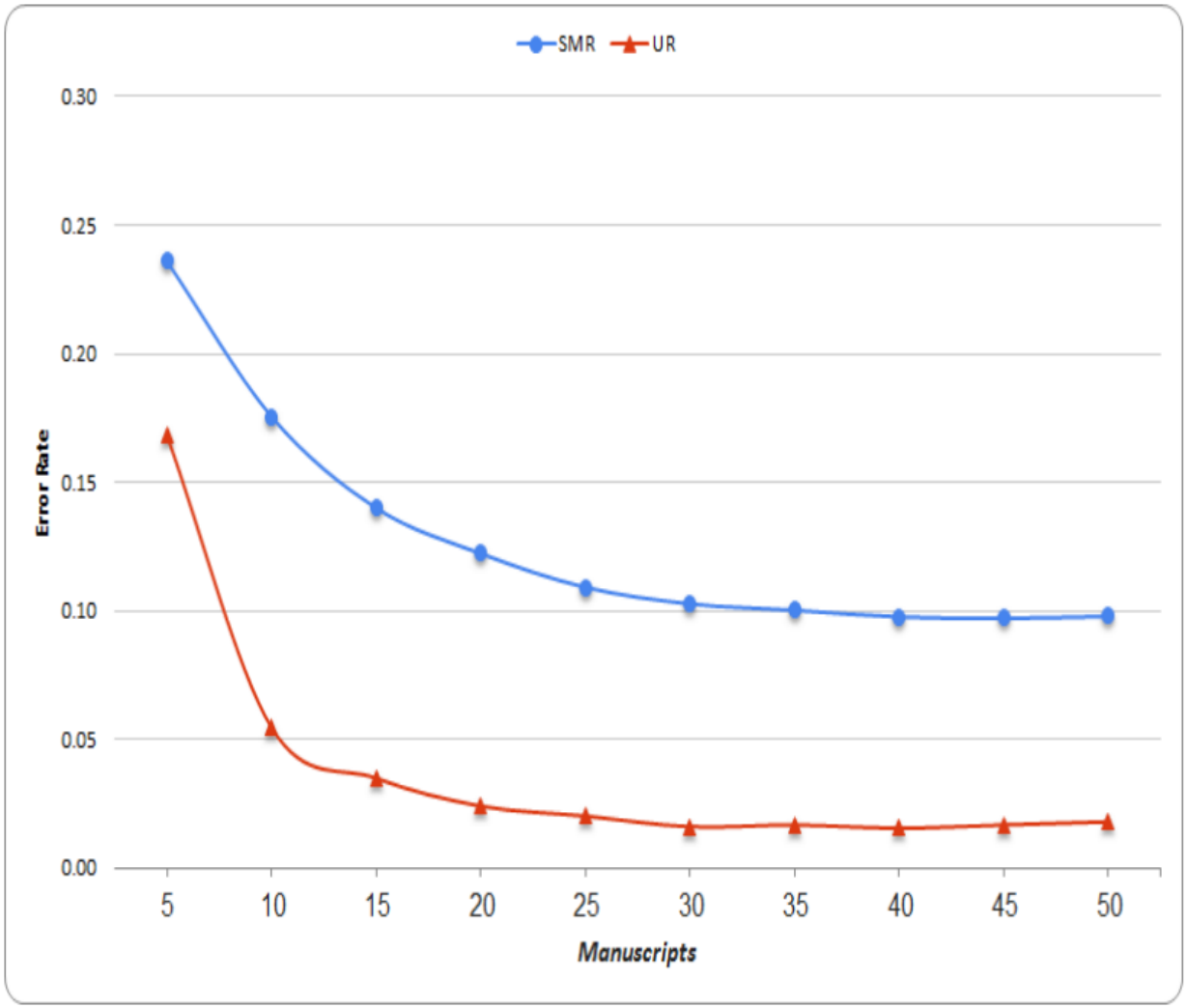
Artificial manuscripts - 2nd Generation

- For this test the authors assume that all manuscripts are copied directly from the original text.
- Each manuscript has reliability p_i chosen from a uniform distribution between 0.20 and 0.99.
- If a word is copied incorrectly, it is randomly replaced by one of k_j possible other words.
- In this way, the authors generate 20 "manuscripts", each with m tokens.

Artificial manuscripts - 2nd Generation



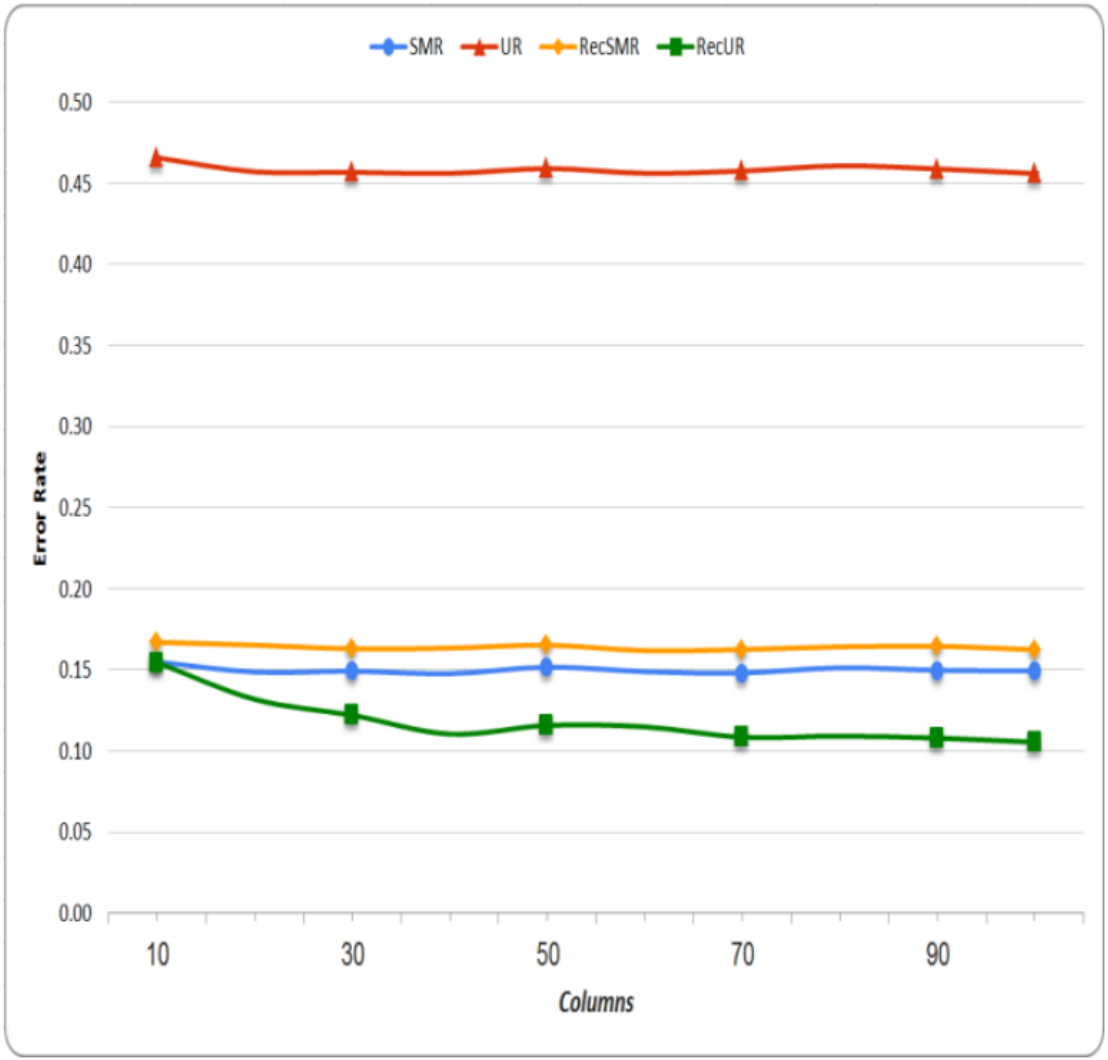
Artificial manuscripts - 2nd Generation



Artificial manuscripts - 3rd Generation

- For this test the authors assume that all manuscripts are copies of copies.
- The authors generate 20 2nd generation manuscripts as before.
- Then they generate 200 3rd generation manuscripts.
- 3rd generation manuscripts are used as input.
- They assume that the clusters are known.

Artificial manuscripts - 3rd Generation



Notre Besoin

- Notre Besoin is an artificial dataset from 2006 created by letting people manually copy an old French manuscript.
- It was used to compare various methods including stemma reconstruction.
- The authors of our paper use it to compare their UR algorithm to stemma reconstruction methods and find no significant difference.

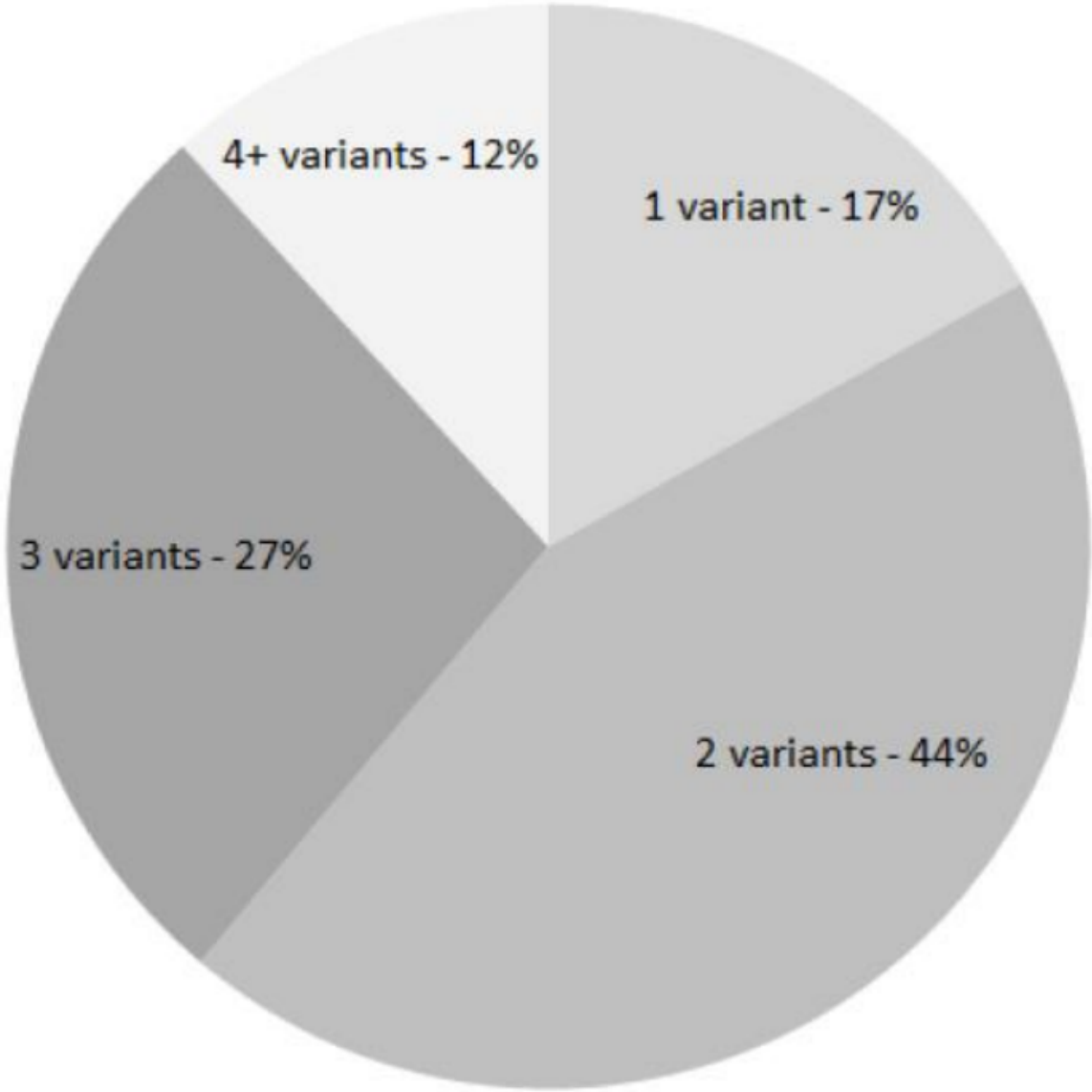
Real world example

- The authors use a synoptic version of a single chapter of the Babylonian Talmud comprising 20 manuscripts and 8564 columns.
- The Manuscripts are split into six clusters by a domain expert.

Real world example

Pre-processing steps:

- Minor spelling-related differences in forms in a column are standardized.
- Merge consecutive columns containing a single token.
- They remain with 5912 columns in the synopsis.


4+ variants - 12%
1 variant - 17%
3 variants - 27%
2 variants - 44%

Real world example

Pre-processing steps:

- Of the 5912 columns, the UR and SMR disagree for 448.
- An expert chooses the most likely correct word between UR and SMR of those 448 columns.
- The expert decides that only 80 columns are significant and resolvable, and UR is better in 82.5%

Conclusion

- The UR algorithm gets better word error rates than SMR on all datasets.
- An expert also judges the UR algorithm as better than SMR in those cases where they didn't agree.

Pros:

- Contribution in a field where little research exists in the context of NLP.

Cons:

- The baseline is weak, there is no extensive comparison to automatic methods or human evaluation.
- Too simple of a contribution since they admit to assuming that:
  - The synopsis and clustering are made by experts by hand.
  - A manuscript has the same probability of being wrong for each word of the manuscript.

Discussion

Any questions or comments?

Questions

1. How could such a method be improved today?
2. How could the evaluation be improved?

Paul Maas. *Textual criticism*. 1958.