A Computational Model of Text Reuse in Ancient Literary Texts

John Lee
Spoken Language Systems
MIT Computer Science and Artificial Intelligence Laboratory
Cambridge, MA 02139, USA
jsylee@csail.mit.edu

Abstract

We propose a computational model of text reuse tailored for ancient literary texts, available to us often only in small and noisy samples. The model takes into account source alternation patterns, so as to be able to align even sentences with low surface similarity. We demonstrate its ability to characterize text reuse in the Greek New Testament.

1 Introduction

Text reuse is the transformation of a source text into a target text in order to serve a different purpose. Past research has addressed a variety of text-reuse applications, including: journalists turning a news agency text into a newspaper story (Clough et al., 2002); editors adapting an encyclopedia entry to an abridged version (Barzilay and Elhadad, 2003); and plagiarizers disguising their sources by removing surface similarities (Uzuner et al., 2005).

A common assumption in the recovery of text reuse is the conservation of some degree of lexical similarity from the *source sentence* to the *derived sentence*. A simple approach, then, is to define a lexical similarity measure and estimate a score threshold; given a sentence in the target text, if the highest-scoring sentence in the source text is above the threshold, then the former is considered to be derived from the latter. Obviously, the effectiveness of this basic approach depends on the degree of lexical similarity: source sentences that are quoted verbatim are easier to identify than those that have been transformed by a skillful plagiarizer.

The crux of the question, therefore, is how to identify source sentences despite their lack of surface similarity to the derived sentences. Ancient literary texts, which are the focus of this paper, present some distinctive challenges in this respect.

1.1 Ancient Literary Texts

"Borrowed material embedded in the flow of a writer's text is a common phenomenon in Antiquity." (van den Hoek, 1996). Ancient writers rarely acknowledged their sources. Due to the scarcity of books, they often needed to quote from memory, resulting in inexact quotations. Furthermore, they combined multiple sources, sometimes inserting new material or substantially paraphrasing their sources to suit their purpose. To compound the noise, the version of the source text available to us today might not be the same as the one originally consulted by the author. Before the age of the printing press, documents were susceptible to corruptions introduced by copyists.

Identifying the sources of ancient texts is useful in many ways. It helps establish their relative dates. It traces the evolution of ideas. The material quoted, left out or altered in a composition provides much insight into the agenda of its author. Among the more frequently quoted ancient books are the gospels in the New Testament. Three of them — the gospels of Matthew, Mark, and Luke — are called the Synoptic Gospels because of the substantial text reuse among them.

Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, pages 472–479,
Prague, Czech Republic, June 2007. ©2007 Association for Computational Linguistics

Target verses (English translation)	Target verses (original Greek)	Source verses (original Greek)
Luke 9:30-33	Luke 9:30-33	Mark 9:4-5
(9:30) **And**, *behold,*	(9:30) **kai** idou	(9:4) kai ōphthē autois
there **talked** *with him two men,*	andres duo **sunelaloun** autō	**Ēlias** sun **Mōusei** kai
which **were Moses** *and* **Elias**.	hoitines **ēsan Mōusēs** kai **Ēlias**	**ēsan sullalountes** tō Iēsou
(9:31) *Who appeared in glory, ...*	(9:31) hoi ophthentes en doxē ...	(no obvious source verse)
(9:32) *But Peter and they that were with him ...*	(9:32) ho de Petros kai hoi sun autō ...	(no obvious source verse)
(9:33) **And** *it came to pass,*	(9:33) **kai** egeneto en tō diachōrizesthai	
as they departed from him,	autous ap' autou eipen **ho Petros**	(9:5) **kai** apokritheis **ho Petros**
Peter *said unto* **Jesus**, *Master,*	pros ton **Iēsoun** epistata	legei tō **Iēsou** rabbi
it is good for us to be here:	**kalon estin hēmas hōde einai**	**kalon estin hēmas hōde einai**
and let us make	**kai poiēsōmen skēnas treis**	**kai poiēsōmen treis skēnas**
three tabernacles; one for thee,	**mian soi kai mian Mōusei**	**soi mian kai Mōusei mian**
and one for Moses, and one for Elias:	**kai mian Ēlia**	**kai Ēlia mian**
not knowing what he said.	mē eidōs ho legei	

Table 1: Luke 9:30-33 and their source verses in the Gospel of Mark. The Greek words with common stems in the target and source verses are bolded. The King James Version English translation is included for reference. §1.2 comments on the text reuse in these verses.

1.2 Synoptic Gospels

The nature of text reuse among the Synoptics spans a wide spectrum. On the one hand, some revered verses, such as the sayings of Jesus or the apostles, were preserved verbatim. Such is the case with Peter's short speech in the second half of Luke 9:33 (see Table 1). On the other hand, unimportant details may be deleted, and new information weaved in from other sources or oral traditions. For example, "Luke often edits the introductions to new sections with the greatest independence" (Taylor, 1972). To complicate matters, it is believed by some researchers that the version of the Gospel of Mark used by Luke was a more primitive version, customarily called *Proto-Mark*, which is no longer extant (Boismard, 1972). Continuing our example in Table 1, verses 9:31-32 have no obvious counterparts in the Gospel of Mark. Some researchers have attributed them to an earlier version of Mark (Boismard, 1972) or to Luke's "redactional tendencies" (Bovon, 2002).

The result is that some verses bear little resemblance to their sources, due to extensive redaction, or to discrepancies between different versions of the source text. In the first case, any surface similarity score alone is unlikely to be effective. In the second, even deep semantic analysis might not suffice.

1.3 Goals

One property of text reuse that has not been explored in past research is *source alternation patterns*. For example, "it is well known that sections of Luke derived from Mark and those of other origins are arranged in continuous blocks" (Cadbury, 1920). This notion can be formalized with features on the blocks and order of the source sentences. The first goal of this paper is to *leverage source alternation patterns to optimize the global text reuse hypothesis.*

Scholars of ancient texts tend to express their analyses qualitatively. We attempt to translate their insights into a quantitative model. To our best knowledge, this is the first sentence-level, quantitative text-reuse model proposed for ancient texts. Our second goal is thus to *bring a quantitative approach to source analysis of ancient texts.*

2 Previous Work

Text reuse is analyzed at the document level in (Clough et al., 2002), which classifies newspaper articles as wholly, partially, or non-derived from a news agency text. The hapax legomena, and sentence alignment based on N-gram overlap, are found to be the most useful features. Considering a document as a whole mitigates the problem of low similarity scores for some of the derived sentences.

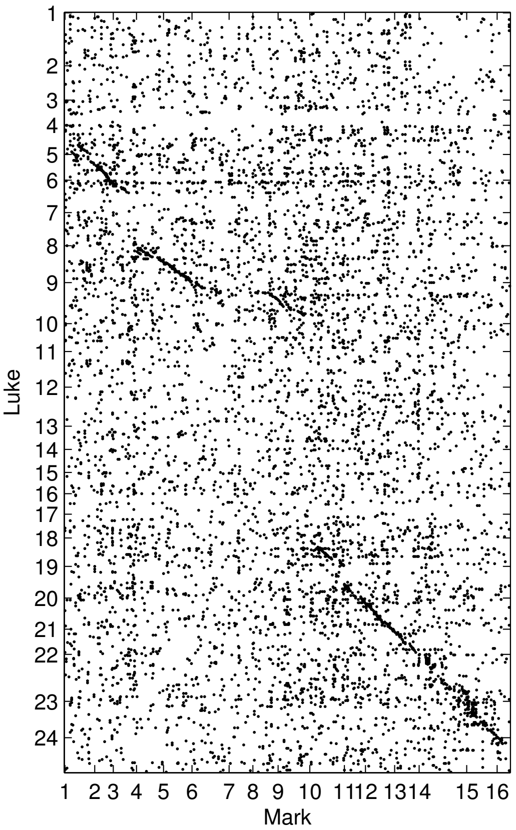

Figure 1: A dot-plot of the cosine similarity measure between the Gospel of Luke and the Gospel of Mark. The number on the axes represent chapters. The thick diagonal lines reflect regions of high lexical similarity between the two gospels.

At the level of short passages or sentences, (Hatzivassiloglou et al., 1999) goes beyond N-gram, taking advantage of WordNet synonyms, as well as ordering and distance between shared words. (Barzilay and Elhadad, 2003) shows that the simple cosine similarity score can be effective when used in conjunction with paragraph clustering. A more detailed comparison with this work follows in §4.2.

In the humanities, reused material in the writings of Plutarch (Helmbold and O'Neil, 1959) and Clement (van den Hoek, 1996) have been manually classified as quotations, reminiscences, references or paraphrases. Studies on the Synoptics have been limited to N-gram overlap, notably (Honoré, 1968) and (Miyake et al., 2004).

Text	Hypothesis	Researcher	Model
L_{train}	$L_{train.B}$	(Bovon, 2002)	*B*
	$L_{train.J}$	(Jeremias, 1966)	*J*
L_{test}	$L_{test.B}$	(Bovon, 2003)	
	$L_{test.J}$	(Jeremias, 1966)	

Table 2: Two models of text reuse of Mark in L_{train} are trained on two different text-reuse hypotheses: The *B* model is on the hypothesis in (Bovon, 2002), and the *J* model, on (Jeremias, 1966). These two models then predict the text-reuse in L_{test}.

3 Data

We assume the Two-Document Theory[1], which hypothesizes that the Gospel of Luke and the Gospel of Matthew have as their common sources two documents: the Gospel of Mark, and a lost text customarily denoted *Q*. In particular, we will consider the Gospel of Luke[2] as the target text, and the Gospel of Mark as the source text.

We use a Greek New Testament corpus prepared by the Center for Computer Analysis of Texts at the University of Pennsylvania[3], based on the text variant from the United Bible Society. The text-reuse hypotheses (i.e., lists of verses deemed to be derived from Mark) of François Bovon (Bovon, 2002; Bovon, 2003) and Joachim Jeremias (Jeremias, 1966) are used. Table 2 presents our notations.

Luke 1:1 to 9:50 (L_{train}, 458 verses) Chapters 1 and 2, narratives of the births of Jesus and John the Baptist, are based on non-Markan sources. Verses 3:1 to 9:50 describe Jesus' activities in Galilee, a substantial part of which is derived from Mark.

Luke Chapters 22 to 24 (L_{test}, 179 verses) These chapters, known as the Passion Narrative, serve as our test text. Markan sources were behind 38% of the verses, according to Bovon, and 7% according to Jeremias.

[1] This theory (Streeter, 1930) is currently accepted by a majority of researchers. It guides our choice of experimental data, but our model does not depend on its validity.

[2] We do not consider the Gospel of Matthew or *Q* in this study. Verses from Luke 9:51 to the end of chapter 21 are also not considered, since their sources are difficult to ascertain (Bovon, 2002).

[3] Obtained through Peter Ballard (personal communication)

4 Approach

For each verse in the target text (a "target verse"), we would like to determine whether it is derived from a verse in the source text (a "source verse") and, if so, which one.

Following the framework of global linear models in (Collins, 2002), we cast this task as learning a mapping F from input verses $\mathbf{x} \in \mathcal{X}$ to a text-reuse hypothesis $\mathbf{y} \in \mathcal{Y} \cup \{\epsilon\}$. $\mathcal{X}$ is the set of verses in the target text. In our case, $\mathbf{x}_{train} = (x_1, \ldots, x_{458})$ is the sequence of verses in L_{train}, and $\mathbf{x}_{test}$ is that of L_{test}. $\mathcal{Y}$ is the set of verses in the source text. Say the sequence $\mathbf{y} = (y_1, \ldots, y_n)$ is the text-reuse hypothesis for $\mathbf{x} = (x_1, \ldots, x_n)$. If y_i is ϵ, then x_i is not derived from the source text; otherwise, y_i is the source verse for x_i. The set of candidates $\mathbf{GEN}(\mathbf{x})$ contains all possible sequences for $\mathbf{y}$, and Θ is the parameter vector. The mapping F is thus:

$$F(\mathbf{x}) = \arg \max_{\mathbf{y} \in \mathbf{GEN}(\mathbf{x})} \Phi(\mathbf{x}, \mathbf{y}) \cdot \Theta$$

4.1 Features

Given the small amount of training data available[4], the feature space must be kept small to avoid overfitting. Starting with the cosine similarity score as the baseline feature, we progressively enrich the model with the following features:

Cosine Similarity [Sim] Treating a target verse as a query to the set of source verses, we compute the cosine similarity, weighted with tf.idf, for each pair of source verse and target verse[5]. This standard bag-of-words approach is appropriate for Greek, a relatively free word-order language. Figure 1 plots this feature on Luke and Mark.

Non-derived verses are assigned a constant score in lieu of the cosine similarity. We will refer to this constant as the *cosine threshold* (C): when the Sim feature alone is used, the constant effectively acts as the threshold above which target verses are considered to be derived. If w_i, w_j are the vectors of words of a target verse and a candidate source verse, then:

$$sim(i,j) = \begin{cases} \frac{w_i \cdot w_j}{\|w_i\| \cdot \|w_j\|} & \text{if derived} \\ C & \text{otherwise} \end{cases}$$

Number of Blocks [Block] Luke can be viewed as alternating between Mark and non-Markan material, and he "prefers to pick up alternatively entire blocks rather than isolated units." (Bovon, 2002) We will use the term *Markan block* to refer to a sequence of verses that are derived from Mark. A verse with a low cosine score, but positioned in the middle of a Markan block, is likely to be derived. Conversely, an isolated verse in the middle of a non-Markan block, even with a high cosine score, is unlikely to be so. The heavier the weight of this feature, the fewer blocks are preferred.

Source Proximity [Prox] When two derived verses are close to one another, their respective source verses are also likely to be close to one another; in other words, derived verses tend to form "continuous blocks" (Cadbury, 1920).

We define *distance* as the number of verses separating two verses. For each pair of consecutive target verses, we take the inverse of the distance between their source verses. This feature is thus intended to discourage a derived verse from being aligned with a source verse that shares some lexical similarities by chance, but is far away from other source verses in the Markan block.

Source Order [Order] "Whenever Luke follows the Markan narrative in his own gospel he follows painstakingly the Markan order", and hence "deviations in the order of the material must therefore be regarded as indications that Luke is not following Mark." (Jeremias, 1966). This feature is a binary function on two consecutive derived verses, indicating whether their source verses are in order. A positive weight for this feature would favor an alignment that respects the order of the source text.

In cases where there are no obvious source verses, such as Luke 9:30-31 in Table 1, the source order

[4] Note that the training set consists of only one $\mathbf{x}_{train}$ — the Gospel of Luke. Luke's only other book, the *Acts of the Apostles*, contains few identifiable reused material.

[5] A targert verse is also allowed to match two consecutive source verses.

and proximity would be disrupted. To mitigate this issue, we allow the Prox and Order features the option of skipping up to two verses within a Markan block in the target text. In our example, Luke 9:30 can skip to 9:32, preserving the source proximity and order between their source verses, Mark 9:4 and 9:5.

Another potential feature is the occurrence of function words characteristic of Luke (Rehkopf, 1959), along the same lines as in the study of the Federalist Papers (Mosteller and Wallace, 1964). These stylistic indicators, however, are unlikely to be as helpful on the sentence level as on the document level. Furthermore, Luke "reworks [his sources] to an extent that, within his entire composition, the sources rarely come to light in their original independent form" (Bovon, 2002). The significance of the presence of these indicators, therefore, is diminished.

4.2 Discussion

This model is both a simplification of and an extension to the one advocated in (Barzilay and Elhadad, 2003). On the one hand, we perform no paragraph clustering or mapping before sentence alignment. Ancient texts are rarely divided into paragraphs, nor are they likely to be large enough for statistical methods on clustering. Instead, we rely on the Prox feature to encourage source verses to stay close to each other in the alignment.

On the other hand, our model makes two extensions to the "Micro Alignment" step in (Barzilay and Elhadad, 2003). First, we add the Block and Prox features to capture source alternation patterns. Second, we place no hard restrictions on the reordering of the source text, opting instead for a soft preference for maintaining the source order through the Order feature. In contrast, deviation from the source order is limited to "flips" between two sentences in (Barzilay and Elhadad, 2003), an assumption that is not valid in the Synoptics[6].

4.3 Evaluation Metric

Our model can make two types of errors: *source error*, when it predicts a non-derived target verse to be derived, or vice versa; and *alignment error*, when it correctly predicts a target verse to be derived, but aligns it to the wrong source verse.

Correspondingly, we interpret the output of our model at two levels: as a binary output, i.e., the target verse is either "derived" or "non-derived"; or, as an alignment of the target verse to a source verse. We measure the precision and recall of the target verses at both levels, yielding two F-measures, F_{source} and F_{align}[7].

Literary dependencies in the Synoptics are typically expressed as pairs of *pericopes* (short, coherent passages), for example, "Luke 22:47-53 // Mark 14:43-52". Likewise, for F_{align}, we consider the output correct if the hypothesized source verse lies within the pericope[8].

5 Experiments

This section presents experiments for evaluating our text-reuse model. §5.1 gives some implementation details. §5.2 describes the training process, which uses text-reuse hypotheses of two different researchers ($L_{train.B}$ and $L_{train.J}$) on the same training text. The two resulting models thus represent two different opinions on how Luke re-used Mark; they then produce two hypotheses on the test text ($\hat{L}_{test.B}$ and $\hat{L}_{test.J}$).

Evaluations of these hypotheses follow. In §5.3, we compare them with the hypotheses of the same two researchers on the test text ($L_{test.B}$ and $L_{test.J}$). In §5.3, we compare them with the hypotheses of seven other representative researchers (Neirynck, 1973). Ideally, when the model is trained on a particular researcher's hypothesis on the train text, its hypothesis on the test text should be closest to the one proposed by the same researcher.

5.1 Implementation

Suppose we align the i^{th} target verse to the k^{th} source verse or to ϵ. Using dynamic programming, their score is the cosine similarity score $sim(i,k)$, added to the best alignment state up to the $(i-1-skip)^{th}$ target verse, where $skip$ can vary from 0 to 2 (see §4.1). If the j^{th} source verse is the aligned

[6] For example, Luke 6:12-19 transposes Mark 3:7-12 and Mark 3:13-19 (Bovon, 2002).

[7] Note that F_{align} is never higher than F_{source} since it penalizes both source and alignment errors.

[8] A more fine-grained metric is individual verse alignment. This is unfortunately difficult to measure. As discussed in §1.2, many derived verses have no clear source verses.

Model	*B*		*J*	
Train Hyp	$L_{train.B}$		$L_{train.J}$	
Metric	F_{source}	F_{align}	F_{source}	F_{align}
`Sim`	0.760	0.646	0.748	0.635
`+Block`	0.961	0.728	0.977	0.743
`All`	**0.985**	**0.949**	**0.983**	**0.936**

Table 3: Performance on the training text, L_{train}. The features are accumulative; `All` refers to the full feature set.

verse in this state, then $score(i, k)$ is:

$$\begin{aligned} sim(i,k) + \max_{j,skip}\{score(i-1-skip, j) \\ +w_{prox} \cdot prox(j,k) + w_{order} \cdot order(j,k) \\ -w_{block} \cdot block(j,k)\} \end{aligned}$$

If both j and k are aligned (i.e., not ϵ), then:

$$\begin{aligned} prox(j,k) &= \frac{1}{dist(j,k)} \\ order(j,k) &= 1 \text{ if } j \geq k \\ block(j,k) &= 1 \text{ if starting new block} \end{aligned}$$

Otherwise these are set to zero.

5.2 Training Results

The model takes only four parameters: the weights for the `Block`, `Prox` and `Order` features, as well as the cosine threshold (C). They are empirically optimized, accurate to 0.01, on the two training hypotheses listed in Table 2, yielding two models, *B* and *J*.

Table 3 shows the increasing accuracy of both models in describing the text reuse in L_{train} as more features are incorporated. The `Block` feature contributes most in predicting the block boundaries, as seen in the jump of F_{source} from `Sim` to `+Block`. The `Prox` and `Order` features substantially improve the alignment, boosting the F_{align} from `+Block` to `All`.

Both models *B* and *J* fit their respective hypotheses to very high degrees. For *B*, the only significant source error occurs in Luke 8:1-4, which are derived verses with low similarity scores. They are transitional verses at the beginning of a Markan block. For *J*, the pericope Luke 6:12-16 is wrongly predicted as derived.

Model	*B*		*J*	
Test Hyp	$L_{test.B}$		$L_{test.J}$	
Metric	F_{source}	F_{align}	F_{source}	F_{align}
`Sim`	0.579	0.382	0.186	0.144
`+Block`	0.671	0.329	0.743	0.400
`All`	**0.779**	**0.565**	**0.839**	**0.839**

Table 5: Performance on the test text, L_{test}.

Most alignment errors are misalignments to a neighboring pericope, typically for verses located near the boundary between two pericopes. Due to their low similarity scores, the model was unable to decide if they belong to the end of the preceding pericope or to the beginning of the following one.

5.3 Test Results

The two models trained in §5.2, *B* and *J*, are intended to capture the characteristics of text reuse in L_{train} according to two different researchers. When applied on the test text, L_{test}, they produce two hypotheses, $\hat{L}_{test.B}$ and $\hat{L}_{test.J}$. Ideally, they should be similar to the hypotheses offered by the same researchers (namely, $L_{test.B}$ and $L_{test.J}$), and dissimilar to those by other researchers. We analyze the first aspect in §5.3, and the second aspect in §5.3.

Comparison with Bovon and Jeremias

Table 4 shows the output of *B* and *J* on L_{test}. As more features are added, their output increasingly resemble $L_{test.B}$ and $L_{test.J}$, as shown in Table 5.

Both $\hat{L}_{test.B}$ and $\hat{L}_{test.J}$ contain the same number of Markan blocks as the "reference" hypotheses proposed by the respective scholars. In both cases, the pericope Luke 22:24-30 is correctly assigned as non-derived, despite their relatively high cosine scores. This illustrates the effect of the `Block` feature.

As for source errors, both *B* and *J* mistakenly assign Luke 22:15-18 as Markan, attracted by the high similarity score of Luke 22:18 with Mark 14:25. *B*, in addition, attributes another pericope to Mark where Bovon does not. Despite the penalty of lower source proximity, it wrongly aligned Luke 23:37-38 to Mark 15:2, misled by a specific title of Jesus that happens to be present in both.

```
Chp 22.....................................................................23.....................................................

Sim xx--x-x-xxxxxx-xxxxx-xx----------x---xxx-x---xx--x-xxx-xxxxxx-xx-x-xxxxx-x--x--xx-----x--xxx--xx------x-x-xxx---xxxx---xxxxx--
All xxxxxxxxxxxxxxxxxx-------------------------------xxxxxxxxxxxxxxxxxxxxxxxxxxxx-----------------------------xxxxxxxxxxxxxxxxx---
Bov xxxxxxxxxxxxxx--------------------------------xxxxxxxxxxxxxxxxxxxxxxxxxxxxxx-------------------------------------xxxxxxxxxxxxx

Sim xx--x-x-xxxxxx-xxxxx-xx----------x---xxx-x---xx--x-xxx-xxxxxx-xx-x-xxxxx-x--x--xx-----x--xxx--xx------x-x-xxx---xxxx---xxxxx--
All xxxxxxxxxxxxxxxxxx------------------------------------------------------------------------------------------------------------
Jer xxxxxxxxxxxxx-----------------------------------------------------------------------------------------------------------------

Gru xxxxxxxxxxxxx-------xxx---------xx-----------------xx--------------------x-----------------------------xx--x-----xx-x---xxx---
Haw xxxxxxxxxxxxx----x---x-------------------x---xx----xxx------x---------x--------------------x---x-------x---------xxx-----xx---
Reh xxxxxxxxxxxxx------x-------------------------------xx------------------------------------------x------------------------------
Snd --------------------xxx---------xx-----------------xxxxxxxxxx-------xxx-------------------------------------------------------
Srm xxxxxxxxxxxxxx------xxx---------xx--------------------------------------------------------------------------------------------
Str xxxxxxxxxxxxx----x---x-------------------x---xx----xxxxxxxxxx---------x--x-----------------x--xx------xx---x-----xxx-----xx---
Tay xxxxxxxxxxxxx--------x-----------x-----------x---x-xxxxxxxxxx------------x---------------------x-------x---x-----xx---xxxxxx--

Chp 24...................................................

Sim xxx---x-xx---------x---x-------xxx-x---x--x-xxx-x---- (Model B Sim)
All ----------------------------------------------------- (Model B All)
Bov xxxxxxxxxxx------------------------------------------ (Bovon)

Sim xxx---x-xx---------x---x-------xxx-x---x--x-xxx-x---- (Model J Sim)
All ----------------------------------------------------- (Model J All)
Jer ----------------------------------------------------- (Jeremias)

Gru -x---x----------------------------------------------- (Grundmann)
Haw -----x----------------------------------------------- (Hawkins)
Reh ----------------------------------------------------- (Rehkopf)
Snd -x---x---x------------------------------------------- (Schneider)
Srm ----------------------------------------------------- (Schürmann)
Str -----x----------------------------------------------- (Streeter)
Tay ---------x------------------------------------------- (Taylor)
```

Table 4: Output of models *B* and *J*, and scholarly hypotheses on the test text, L_{test}. The symbol '`x`' indicates that the verse is derived from Mark, and '`–`' indicates that it is not. The hypothesis from (Bovon, 2003), labelled '`Bov`', is compared with the `Sim` (baseline) output and the `All` output of model *B*, as detailed in Table 5. The hypothesis from (Jeremias, 1966), '`Jer`', is similarly compared with outputs of model *J*. Seven other scholarly hypotheses are also listed.

Elsewhere, *B* is more conservative than Bovon in proposing Markan derivation. For instance, the pericope Luke 24:1-11 is deemed non-derived, an opinion (partially) shared by some of the other seven researchers.

Comparison with Other Hypotheses

Another way of evaluating the output of *B* and *J* is to compare them with the hypotheses of other researchers. As shown in Table 6, $\hat{L}_{test.B}$ is more similar to $L_{test.B}$ than to the hypothesis of other researchers[9]. In other words, when the model is trained on Bovon's text-reuse hypothesis on the train text, its prediction on the test text matches most closely with that of the same researcher, Bovon.

Hypothesis	*B* ($\hat{L}_{test.B}$)	*J* ($\hat{L}_{test.J}$)
Bovon ($L_{test.B}$)	**0.838**	0.676
Jeremias ($L_{test.J}$)	0.721	**0.972**
Grundmann	0.726	0.866
Hawkins	0.737	0.877
Rehkopf	0.721	0.950
Schneider	0.676	0.782
Schürmann	0.698	0.950
Streeter	0.771	0.821
Taylor	0.793	0.821

Table 6: Comparison of the output of the models *B* and *J* with hypotheses by prominent researchers listed in (Neirynck, 1973). The metric is the percentage of verses deemed by both hypotheses to be "derived", or "non-derived".

[9] This is the list of researchers whose opinions on L_{test} are considered representative by (Neirynck, 1973). We have simplified their hypotheses, considering those "partially assimilated" and "reflect the influence of Mark" to be non-derived from Mark.

The differences between Bovon and the next two most similar hypotheses, Taylor and Streeter, are not statistically significant according to McNemar's test ($p = 0.27$ and $p = 0.10$ respectively), possibly a reflection of the small size of L_{test}; the differences are significant, however, with all other hypotheses ($p < 0.05$). Similar results are observed for Jeremias and $\hat{L}_{test.J}$.

6 Conclusion & Future Work

We have proposed a text-reuse model for ancient literary texts, with novel features that account for source alternation patterns. These features were validated on the Lukan Passion Narrative, an instance of text reuse in the Greek New Testament.

The model's predictions on this passage are compared to nine scholarly hypotheses. When tuned on the text-reuse hypothesis of a certain researcher on the train text, it favors the hypothesis of the same person on the test text. This demonstrates the model's ability to capture the researcher's particular understanding of text reuse.

While a computational model alone is unlikely to provide definitive answers, it can serve as a supplement to linguistic and literary-critical approaches to text-reuse analysis, and can be especially helpful when dealing with a large amount of candidate source texts.

Acknowledgements

This work grew out of a term project in the course "Gospel of Luke", taught by Professor François Bovon at Harvard Divinity School. It has also benefited much from discussions with Dr. Steven Lulich.

References

R. Barzilay and N. Elhadad. 2003. *Sentence Alignment for Monolingual Comparable Corpora*. Proc. EMNLP.

M. E. Boismard. 1972. *Synopse des quatre Evangiles en français, Tome II.* Editions du Cerf, Paris, France.

F. Bovon. 2002. *Luke I: A Commentary on the Gospel of Luke 1:1-9:50*. Hermeneia. Fortress Press. Minneapolis, MN.

F. Bovon. 2003. The Lukan Story of the Passion of Jesus (Luke 22-23). *Studies in Early Christianity*. Baker Academic, Grand Rapids, MI.

H. J. Cadbury. 1920. *The Style and Literary Method of Luke*. Harvard Theological Studies, Number VI. George F. Moore and James H. Ropes and Kirsopp Lake (ed). Harvard University Press, Cambridge, MA.

P. Clough, R. Gaizauskas, S. S. L. Piao and Y. Wilks. 2002. *METER: MEasuring TExt Reuse*. Proc. ACL.

M. Collins. 2002. *Discriminative Training Methods for Hidden Markov Models: Theory and Experiments with Perceptron Algorithms.* Proc. EMNLP.

V. Hatzivassiloglou, J. L. Klavans and E. Eskin. 1999. *Detecting Text Similarity over Short Passages: Exploring Linguistic Feature Combinations via Machine Learning*. Proc. EMNLP.

W. C. Helmbold and E. N. O'Neil. 1959. *Plutarch's Quotations*. Philological Monographs XIX, American Philological Association.

A. M. Honoré. 1968. *A Statistical Study of the Synoptic Problem.* Novum Testamentum, Vol. 10, p.95-147.

J. Jeremias. 1966. *The Eucharistic Words of Jesus*. Scribner's, New York, NY.

M. Miyake, H. Akama, M. Sato, M. Nakagawa and N. Makoshi. 2004. *Tele-Synopsis for Biblical Research: Development of NLP based Synoptic Software for Text Analysis as a Mediator of Educational Technology and Knowledge Discovery*. Proc. IEEE International Conference on Advanced Learning Technologies (ICALT).

F. Mosteller and D. L. Wallace. 1964. *Inference and Disputed Authorship: The Federalist.* Addison Wesley, Reading, MA.

F. Neirynck. 1973. *La matière marcienne dans l'évangile de Luc*. L'Évangile de Luc, Problèmes littéraires et théologiques. Editions Duculot, Belgium.

F. Rehkopf. 1959. *Die lukanische Sonderquelle*. Wissenschaftliche Untersuchungen zum Neuen Testament, Vol. 5. Tübingen, Germany.

B. H. Streeter. 1930. *The Four Gospels: A Study of Origins*. MacMillan. London, England.

V. Taylor. 1972. *The Passion Narrative of St. Luke: A Critical and Historical Investigation*. Society for New Testament Studies Monograph Series, Vol. 19. Cambridge University Press, Cambridge, England.

O. Uzuner, B. Katz and T. Nahnsen. 2005. *Using Syntactic Information to Identify Plagiarism*. Proc. 2nd Workshop on Building Educational Applications using NLP. Ann Arbor, MI.

A. van den Hoek. 1996. *Techniques of Quotation in Clement of Alexandria — A View of Ancient Literary Working Methods*. Vigiliae Christianae, Vol 50, p.223-243. E. J. Brill, Leiden, The Netherlands.