Supervised Sentence Fusion with Single-Stage Inference

Kapil Thadani & Kathy McKeown

Columbia University
The heavy-metal group Metallica filed a federal lawsuit in 2000 against Napster for copyright infringement, charging that Napster encouraged users to trade copyrighted material without the band’s permission.

The heavy metal rock band Metallica, rap artist Dr. Dre and the RIAA have sued Napster, developer of Internet sharing software, alleging the software enables the acquisition of copyrighted music without permission.

The heavy-metal band Metallica sued Napster and three universities for copyright infringement and racketeering, seeking $10 million in damages.

Metallica sued Napster for copyright infringement
The heavy-metal group Metallica filed a federal lawsuit in 2000 against Napster for copyright infringement, charging that Napster encouraged users to trade copyrighted material without the band’s permission.

The heavy metal rock band Metallica, rap artist Dr. Dre and the RIAA have sued Napster, developer of Internet sharing software, alleging the software enables the acquisition of copyrighted music without permission.

The heavy-metal band Metallica sued Napster and three universities for copyright infringement and racketeering, seeking $10 million in damages.

Metallica sued Napster for copyright infringement
The heavy-metal group Metallica filed a federal lawsuit in 2000 against Napster for copyright infringement, charging that Napster encouraged users to trade copyrighted material without the band’s permission.

The heavy metal rock band Metallica, rap artist Dr. Dre and the RIAA have sued Napster, developer of Internet sharing software, alleging the software enables the acquisition of copyrighted music without permission.

The heavy-metal band Metallica sued Napster and three universities for copyright infringement and racketeering, seeking $10 million in damages.

Metallica sued Napster for copyright infringement
The heavy-metal group Metallica filed a federal lawsuit in 2000 against Napster for copyright infringement, charging that Napster encouraged users to trade copyrighted material without the band’s permission.

The heavy metal rock band Metallica, rap artist Dr. Dre and the RIAA have sued Napster, developer of Internet sharing software, alleging the software enables the acquisition of copyrighted music without permission.

The heavy-metal band Metallica sued Napster and three universities for copyright infringement and racketeering, seeking $10 million in damages.

Metallica sued Napster for copyright infringement
sentence fusion

definition

- merge **two or more** sentences to produce a single sentence
- preserve **salient** information
sentence fusion

definition

- merge **two or more** sentences to produce a single sentence
- preserve **salient** information

**annotation**  
Daumé III & Marcu (2004), Marsi & Krahmer (2005), Krahmer et al. (2008), McKeown et al. (2010)

**unsupervised**  
Barzilay & McKeown (2005), Filippova & Strube (2008), Filippova (2010), Thadani & McKeown (2011), Boudin & Morin (2013)

**supervised**  
Elsner & Santhanam (2011)
sentence fusion

definition

- merge **exactly two** sentences to produce a single sentence
- preserve **salient** information

annotation: Daumé III & Marcu (2004), Marsi & Krahmer (2005), Krahmer et al. (2008), McKeown et al. (2010)

unsupervised: Barzilay & McKeown (2005), Filippova & Strube (2008), Filippova (2010), Thadani & McKeown (2011), Boudin & Morin (2013)

supervised: Elsner & Santhanam (2011)
sentence fusion

definition

- merge **exactly two** sentences to produce a single sentence
- preserve **only repeated** information

annotation

Daumé III & Marcu (2004), Marsi & Krahmer (2005),
Krahmer et al. (2008), McKeown et al. (2010)

unsupervised

Barzilay & McKeown (2005), Filippova & Strube (2008),
Filippova (2010), Thadani & McKeown (2011), Boudin & Morin (2013)

supervised

Elsner & Santhanam (2011)
sentence fusion

definition

- merge **two or more** sentences to produce a single sentence
- preserve **salient** information

**annotation**
- Daumé III & Marcu (2004), Marsi & Krahmer (2005), Krahmer et al. (2008), McKeown et al. (2010)

**unsupervised**
- Barzilay & McKeown (2005), Filippova & Strube (2008), Filippova (2010), Thadani & McKeown (2011), Boudin & Morin (2013)

**supervised**
- Elsner & Santhanam (2011)
sentence fusion

definition

▶ merge **two or more** sentences to produce a single sentence
▶ preserve **salient** information

challenges

no standard dataset for learning and evaluation
   - Elsner & Santhanam (2011) dataset es11 can’t be distributed
   - McKeown et al. (2010) dataset mrtm10 noisy for intersections

difficult annotation task
   - Daumé & Marcu (2004), Krahmer et al. (2008), McKeown et al. (2010)
   - would prefer **natural** data
sentence fusion
this talk

new corpus of \{2, 3, 4\}-way fusions
+ **large**: \(\sim 2000\) instances; 6 times larger than ES11 and MRTM10
+ **natural**: derived from summary evaluation annotations
+ **available**: raw data distributed by NIST

new inference approach for supervised fusion
+ **optimal**: always finds highest scoring fusion
+ **holistic**: jointly identifies salient words and linearizes sentence
+ **expressive**: permits rich features and lexical constraints
outline

▶ overview

▶ corpus construction

▶ supervised fusion approach

▶ experiments
fusion corpus

data source

pyramid evaluation of summaries (Nenkova et al., 2007)
  ▶ DUC 2005–2007, TAC 2008–2011

for a group of human summaries on a particular news topic, annotators have identified:
  i **SCUs**: “semantic content units” — atomic units of information
  ii **SCU contributors**: summary text that expresses SCU

for an SCU with >1 contributors, we map:
  ▶ summary sentences → input sentences for fusion
  ▶ SCU label → gold fusion output
The heavy-metal group Metallica filed a federal lawsuit in 2000 against Napster for copyright infringement, charging that Napster encouraged users to trade copyrighted material without the band’s permission.

The heavy metal rock band Metallica, rap artist Dr. Dre and the RIAA have sued Napster, developer of Internet sharing software, alleging the software enables the acquisition of copyrighted music without permission.

The heavy-metal band Metallica sued Napster and three universities for copyright infringement and racketeering, seeking $10 million in damages.

Metallica sued Napster for copyright infringement
fusion corpus

filtering

only keep SCUs when:

1. the SCU seems to address main concept of source sentences
2. the label is a complete sentence
3. label words come from the source sentences

after filtering: 1858 fusion instances

- 2-way: 873
- 3-way: 569
- 4-way: 416
fusion corpus
download

pyramid data available from NIST
  ▶ duc.nist.gov & nist.gov/tac
fusion corpus
download

pyramid data available from NIST
▶ duc.nist.gov & nist.gov/tac
outline

▶ overview

▶ corpus construction

▶ supervised fusion approach

▶ experiments
supervised fusion approach

“single stage” inference

most previous work has 2-3 stages

1. align input sentences
2. select output content using a dependency graph
3. linearize tree using LM and heuristics

dthis work:

- based on a new supervised approach for sentence compression
  (Thadani & McKeown, CoNLL 2013)
- ILP to optimally recover content and ordering
- implicit alignment via redundancy features and constraints
supervised fusion approach

inference

\[ \hat{F} = \arg \max_{F} \text{score}(F) \]

\[ = \arg \max_{F} \mathbf{w}^{\top} \Phi(F) \]
supervised fusion approach

inference

\[ \hat{F} = \arg \max_{x,y} \sum_i x_i \cdot w_{\text{tok}}^\top \phi(t_i) + \sum_{i,j} y_{ij} \cdot w_{\text{ngr}}^\top \phi(\langle t_i, t_j \rangle) \]
supervised fusion approach

\[ \hat{F} = \arg \max_{x,y} \sum_i x_i \cdot w_{tok} \phi(t_i) + \sum_{i,j} y_{ij} \cdot w_{ngr} \phi(\langle t_i, t_j \rangle) \]

token score

ngram score
supervised fusion approach

\[ \hat{F} = \arg \max_{x, y} \sum_i x_i \cdot w_{tok}^{\top} \phi(t_i) + \sum_{i,j} y_{ij} \cdot w_{ngr}^{\top} \phi(\langle t_i, t_j \rangle) \]

inference

token score

ngram score

indicator variables

The heavy-metal group Metallica filed a federal lawsuit in 2000 against ...

The heavy metal rock band Metallica, rap artist Dr. Dre and the RIAA ...
supervised fusion approach

inference

\[
\hat{F} = \arg \max_{x, y} \sum_i x_i \cdot w_{\text{tok}}^T \phi(t_i) + \sum_{i,j} y_{ij} \cdot w_{\text{ngr}}^T \phi(\langle t_i, t_j \rangle)
\]

features

- **salience**: contextual POS patterns and morphological features
- **fluency**: LM score, POS + dependency features for n-gram
- **fidelity**: whether n-gram is in the input
- **pseudo-normalization**: to account for length variation
supervised fusion approach

inference

\[ \hat{F} = \arg \max_{x,y} \sum_i x_i \cdot w_{\text{tok}}^T \phi(t_i) + \sum_{i,j} y_{ij} \cdot w_{\text{ngr}}^T \phi(\langle t_i, t_j \rangle) \]

token score

ngram score

learned parameters

- structured perceptron with averaging (Collins, 2002)
- with minibatches (Zhao & Huang, 2013)
supervised fusion approach
ILP constraints

- selected tokens $x$ and n-grams $y$ are consistent
  - $y_{ij}$ activates $x_i$ and $x_j$
  - $x_i$ activates exactly one $y_{i*}$ and $y_{*i}$

- $y$ forms an acyclic, connected path
supervised fusion approach
commodity flow variables + constraints

- commodity carried in real-valued variables between all pairs of tokens
  \[ y_{ij} \quad \text{iff} \quad x_i > 0 \Rightarrow \text{consistent with n-gram variables} \]

- active tokens consume 1 unit of commodity
  \[ n \quad \text{⇒} \quad \text{prevents cycles} \]

- originate at a single point (START)
  \[ \Rightarrow \text{guarantees connectivity} \]
supervised fusion approach
commodity flow backbone for n-grams

...group Metallica filed a federal lawsuit in 2000 against Napster for copyright infringement ...

... rock band Metallica, rap artist Dr. Dre and the RIAA have sued Napster, developer of ...
supervised fusion approach
example: redundancy as salience

$S_1$ The heavy-metal group Metallica filed a federal lawsuit in 2000 against Napster for copyright infringement, charging that Napster encouraged users to trade copyrighted material without the band’s permission.

$S_2$ The heavy metal rock band Metallica, rap artist Dr. Dre and the RIAA have sued Napster, developer of Internet sharing software, alleging the software enables the acquisition of copyrighted music without permission.

$S_3$ The heavy-metal band Metallica sued Napster and three universities for copyright infringement and racketeering, seeking $10 million in damages.

FUSION Metallica sued Napster for copyright infringement
supervised fusion approach
exploiting redundancy

want to recognize input redundancy
▶ identify synonym groups across sentences for NN*, VB*, JJ*, RB*
e.g., {Metallica}, {band, group}, {charging, alleging}
▶ support features: how many sentences does the group for a token appear in, conjoined with POS class

want to avoid output redundancy
▶ “Metallica and Metallica sued Napster and ...”
▶ redundancy constraints: each group must appear no more than once in the output
outline

▶ overview

▶ corpus construction

▶ supervised fusion approach

▶ experiments
experiments

systems

compression
▶ state-of-the-art for sentence compression (Thadani & McKeown, CoNLL 2013)
▶ no support features or redundancy constraints

compression over contributors
▶ human-annotated spans that capture the SCU concept in the source
▶ strong baseline: 35% of SCU labels exactly match a contributor

this work: compression + support features + redundancy constraints
experiments
n-gram overlap

![Graph showing n-gram overlap with lines for compression, + contributors, and this work.](image)
experiments

informativeness

| content words                  | P%  | R%  | F₁% |
|-------------------------------|-----|-----|-----|
| compression                   | 40.05 | 28.20 | 30.17 |
| + only contributors           | 55.27† | 36.79 | 39.95 |
| this work                     | 49.01 | 45.09† | 44.42† |

**bold** significant vs others under Wilcoxon’s signed rank test

† significant vs others under paired t-test

content words (nouns + verbs) useful for informativeness in compression (Hori & Furui, 2004)
experiments
grammaticality

| syntactic rels | F$_1$% | Stanford | RASP |
|---------------|-------|----------|------|
| compression   | 14.19 | 12.71    |      |
| + only contributors | **22.81**† | **20.24**† |      |
| this work     | 22.81† | 21.25†   |      |

**bold** significant vs others under Wilcoxon’s signed rank test
† significant vs others under paired t-test

RASP F% correlates with human judgments of fluency in compression
(Napoles & Callison-Burch, 2011)
Elian returned to Cuba on June 28, 2000.

After a final appeal by the Miami relatives was denied and the court order blocking his return expired, Elian returned with his father to Cuba on June 28, 2000.

On June 28, the Supreme Court rejected a final appeal; Elian returned home to Cuba, was celebrated in the media and returned to his home and schooling.

gold SCU Elian returned with his father to Cuba on June 28, 2000

compression Elian returned to Cuba on June returned with his father rejected a final appeal

+ contribs Elian returned to home to Cuba

this work Elian returned to Cuba on June 28
Jennings, who quit smoking several years ago, will undergo chemotherapy in New York.

ABC announced that Jennings would continue to anchor the news during chemotherapy treatment, but he was unable to do so.

Peter Jennings hoarsely announced he had lung cancer on April 5, 2005 and would begin outpatient chemotherapy in New York.

Jennings will undergo chemotherapy in New York.

ABC announced that 2005.

would begin outpatient chemotherapy chemotherapy treatment.

ABC announced that Jennings would undergo chemotherapy in New York.
new corpus of natural fusions
  ▶ large enough for supervised learning
  ▶ available to all (once NIST is back online)

optimal inference approach for supervised fusion
  ▶ avoids hard alignment, content selection
  ▶ soft support features + redundancy

future work
  ▶ joint inference with rich syntactic structure
</talk>