Plain English Summarization of Contracts

Workshop on Natural Legal Language Processing (NLLP)

Laura Manor and Jessy Li
Department of Linguistics, The University of Texas at Austin
By using our Services, you are agreeing to these Terms, our Trainer Guidelines, and our Privacy Policy. If you are the parent or legal guardian of a child under the age of 13 (the “Parent”), you are agreeing to these Terms on behalf of yourself and your child(ren) who are authorized to use the Services pursuant to these Terms and in our Privacy Policy. If you don’t agree to these Terms, our Trainer Guidelines, and our Privacy Policy, do not use the Services.
98% users do not read Terms of Service (TOS)

Users did not notice clauses in an app TOS:

• allowing the company to share data with the user’s employer and the NSA
• requiring the user’s first-born child as payment for access

Why?

Obar and Oeldorf-Hirsch (2018)
1. Information overload
   “too long, too numerous, and taking up too much time”

2. Nothing to hide
   “the individual in question perceives that policies are irrelevant because the individual is doing nothing wrong, companies will not bother them, and only those who are breaking the rules are affected”

3. Difficult to understand
   “individuals perceive that they are unable to understand the language in TOS and privacy policies”

Obar and Oeldorf-Hirsch (2018)
1. Information overload  
   “too long, too numerous, and taking up too much time”

2. Nothing to hide  
   “the individual in question perceives that policies are irrelevant because the individual is doing nothing wrong, companies will not bother them, and only those who are breaking the rules are affected”

3. Difficult to understand  
   “individuals perceive that they are unable to understand the language in TOS and privacy policies”

Obar and Oeldorf-Hirsch (2018)
By using our Services, you are agreeing to these Terms, our Trainer Guidelines, and our Privacy Policy. If you are the parent or legal guardian of a child under the age of 13 (the “Parent”), you are agreeing to these Terms on behalf of yourself and your child(ren) who are authorized to use the Services pursuant to these Terms and in our Privacy Policy. If you don’t agree to these Terms, our Trainer Guidelines, and our Privacy Policy, do not use the Services.

Original: 78 words

By playing this game, you agree to these terms. If you're under 13 and playing, your parent/guardian agrees on your behalf.

Summary: 19 words
Summarization in Plain English

Building a dataset

Characteristics of the dataset

Are we there yet?
Summarization in Plain English

Building a dataset

Characteristics of the dataset

Are we there yet?
Single Document Summarization

Goal: produce a shortened summary of a document that captures its main semantic content

**Abstractive:**
the reference summaries include novel sentences and phrases

**Extractive:**
the reference summaries include sentences from the original text

How do we know automated techniques work?
Single Document Summarization

1. Gather sizable corpus of parallel texts and reference summaries
2. Generate summaries
3. Compare generated summaries and reference summaries using a similarity metric, typically ROUGE
What is “Plain English”

• Makes a text accessible for as many English speakers as possible.
• Guidelines for plain English suggest:
  – use words with Saxon etymologies rather than Latin/Romance etymologies
  – use short words, sentences, and paragraphs, etc.

(Mellinkoff, 2004; Tiersma, 2000)
Existing corpora in

- **news** (Over et al., 2007; See et al., 2017; Grusky et al., 2018)
- **scientific writing** (TAC, 2014; Jaidka et al., 2016; Yasunaga et al., 2019),
- **legal case reports** (Gal-gani et al., 2012)
- etc.

We present the first dataset in the genre of contract summarization in plain English.
Summarization in Plain English

Building a dataset

Characteristics of the dataset

Are we there yet?
“I have read and agree to the Terms” is the biggest lie on the web. We aim to fix that.
TL;DRLegal

84 sets sourced from 9 documents:

- Pokemon GO Terms of Service,
- TLDRLegal Terms of Service,
- Minecraft End User Licence Agreement,
- YouTube Terms of Service,
- Android SDK License Agreement (June 2014),
- Google Play Game Services (May 15th, 2013),
- Facebook Terms of Service (Statement of Rights and Responsibilities),
- Dropbox Terms of Service, and
- Apple Website Terms of Service.
Our goal is to deliver advertising; and other commercial or sponsored content that is valuable to our users and advertisers. In order to help us do that, you agree to the following: You give us permission to use your name, profile picture, content, and information in connection with commercial, sponsored, or related content (such as a brand you like) served or enhanced by us. This means, for example, that you permit a business or other entity to pay us to display your name and/or profile picture with your content or information, without any compensation to you. If you have selected a specific audience for your content or information, we will respect your choice when we use it. We do not give your content or information to advertisers without your consent. You understand that we may not always identify paid services and communications as such.

Reference Summary: Facebook can use any of your stuff for any reason they want without paying you, for advertising in particular.
TOS;DR

421 total sets of parallel text sourced from 166 documents by 122 companies

- Sets consisted of section of an agreement text and 1-3 human-written summaries
- Not all summaries were quality summaries; annotation was necessary
- The annotator selected the most informative summary to be used in this paper
- 361 sets had at least one quality summary in the set
**Original Text:** When you upload, submit, store, send or receive content to or through our Services, you give Google (and those we work with) a worldwide license to use, host, store, reproduce, modify, create derivative works (such as those resulting from translations, adaptations or other changes we make so that your content works better with our Services), communicate, publish, publicly perform, publicly display and distribute such content.

**Reference Summary:** The copyright license you grant is “for the limited purpose of operating, promoting, and improving” existing and new Google Services. However, please note that the license does not end if you stop using the Google services.

**Summary 2:** The copyright license that users grant this service is limited to the parties that make up the service’s broader platform.

**Summary 3:** Limited copyright license to operate and improve all Google Services.
Summarization in Plain English

Building a dataset

Characteristics of the dataset

Are we there yet?
Our Dataset

446 sets of pairs of texts with human-written summaries

**Abstractive:**
the reference summaries include novel sentences and phrases

**Compressive:**
the reference summaries are shorter than the original text

**Simplified:**
the language in the reference summary is simpler than the original text
Abstractive

![Bar chart showing unique n-grams comparison between DUC2002 and Ours.]

- **1-grams**: DUC2002: 12%, Ours: 41%
- **2-grams**: DUC2002: 45%, Ours: 78%
- **3-grams**: DUC2002: 64%, Ours: 88%
- **4-grams**: DUC2002: 73%, Ours: 92%
Compressive

<table>
<thead>
<tr>
<th></th>
<th>Words</th>
<th>Sentences</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original Text</td>
<td>105.6</td>
<td>3.6</td>
</tr>
<tr>
<td>Reference Summary</td>
<td>17.2</td>
<td>1.2</td>
</tr>
</tbody>
</table>
Simplified

Readability Metrics:
- Flesch-Kincaid formula (Kincaid et al., 1975)
- Coleman-Liau index (Coleman and Liau, 1975)
- SMOG (McLaughlin, 1969)
- Automated readability index (Senter and Smith, 1967)

Original Text: 19.29 years
Reference Summary: 13.29 years
Difference: 6.00 years
Summarization in Plain English

Building a dataset

Characteristics of the dataset

Are we there yet?
Summarization Baselines

**Abstractive:**
generate novel language but, require training data

**Extractive:**
some popular methods do not require training data
### Difference: ~20 ROUGE

<table>
<thead>
<tr>
<th></th>
<th>Ours</th>
<th>DUC 2002</th>
</tr>
</thead>
<tbody>
<tr>
<td>TextRank</td>
<td>17.1</td>
<td>36.7</td>
</tr>
<tr>
<td>KLSum</td>
<td>16.93</td>
<td>35.85</td>
</tr>
<tr>
<td>Lead-K</td>
<td>17.63</td>
<td>39.49</td>
</tr>
<tr>
<td>Random-K</td>
<td>16.09</td>
<td>31.91</td>
</tr>
</tbody>
</table>

**Difference:** 1.54 7.58

**TextRank/KL-Sum:**
- common strong baseline extractive summarization techniques
- select sentences according to some salience rating/score
- usually follows the distribution of words in the original documents
Qualitative Analysis

Good Performance
reference summary is an extract or a compressed version of the original text

TextRank, Lead-K

no sale of personal information. [librarything will not sell or give personally identifiable information to any third party.]

Reference Summary

[librarything will not sell or give personally identifiable information to any third party.]
Qualitative Analysis

Bad Performance
the language in the original text is more complex than the reference summary

TextRank, KLSum, Lead-K

if you provide a submission whether by email or otherwise you agree that it is non confidential unless couchsurfing states otherwise in writing and shall become the sole property of couchsurfing

Reference Summary

if you offer suggestions to the service they become the owner of the ideas that you give them
the original text consists of long sentence and the reference summary is able to abstract over a large portion of the original text

Original Text

subject to your compliance with these terms niantic grants you a limited nonexclusive nontransferable non sublicensable license to download and install a copy of the app on a mobile device and to run such copy of the app solely for your own personal noncommercial purposes. [...] by using the app you represent and warrant that i you are not located in a country that is subject to a u s government embargo or that has been designated by the u s government as a terrorist supporting country and ii you are not listed on any u s government list of prohibited or restricted parties.

Reference Summary

don t copy modify resell distribute or reverse engineer this app.

KLSum

by way of example and not as a limitation you agree that when using the services and content you will not defame abuse harass harm stalk threaten or otherwise violate the legal rights including the rights of privacy and publicity of others [...] or regulation or enable any other individual to do any of the foregoing.
The reference summary uses much shorter sentences and a more informal style

**Original Text**

by way of example and not as a limitation you agree that when using the services and content you will not defame abuse harass harm stalk threaten or otherwise violate the legal rights including the rights of privacy and publicity of others [...] lease the app or your account collect or store any personally identifiable information from the services from other users of the services without their express permission violate any applicable law or regulation or enable any other individual to do any of the foregoing.

**Reference Summary**

don’t be a jerk. don’t hack or cheat. we don’t have to ban you but we can. we’ll also cooperate with law enforcement
Summarization in Plain English

Building a dataset

Characteristics of the dataset

Are we there yet? No
Thank you!

The data set can be found at:
github.com/lauramanor/legal_summarization

manor@utexas.edu
Laws that encourage transparency

• The European Union's GDPR (2018)
  https://eugdpr.org

• The United States' Plain Writing Act (2010)
  https://plainlanguage.gov/

• New York State's Plain English Act (1978)
  https://www.nysenate.gov/legislation/laws/GOB/5-702
arise, unless, receive, whether, example, signal, b, technology, identifier, expressly, transmit, visit, perform, search, partner, understand, conduct, server, child, support, regulation, base, similar, purchase, automatically, mobile, agent, derivative, either, commercial, reasonable, cause, functionality, advertiser, act, ii, thereof, arbitrator, attorney, modification, locate, c, individual, form, following, accordance, hereby, cookie, apps, advertisement

fingerprint, fit, header, targeted, involve, pixel, advance, quality, track, want, stuff, even, guarantee, maintain, beacon, ban, month, prohibit, allow, defend, notification, ownership, acceptance, delete, user, prior, reason, hold, notify, govern, keep, class, change, might, illegal, old, harmless, indemnify, see, assume, deletion, waive, stop, operate, year, enforce, target, many, constitute, posting

Reading Level (Original Texts - Reference Summaries) = 5.89

Flesch-Kincaid formula (Kincaid et al., 1975)
Automated readability index (Senter and Smith, 1967)
mean compression rate: 0.31 (std 0.23)
• Propose the task of summarizing legal documents in plain English
• Present an initial evaluation dataset for this task.
  – online sources dedicated to explaining sections of contracts in plain English
  – manually verify the quality of the summaries
  – We show that our dataset is
    • highly abstractive
    • the summaries are much simpler to read.
• This task is challenging
  – popular unsupervised extractive summarization methods do not perform well on this dataset and,
  – current methods that address the change in register are mostly supervised as well.
• We call for the development of resources for unsupervised simplification and style transfer in this domain.