DARPA: Semantic Textual Similarity Workshop
March 13, 2012

Sameer S Pradhan
Raytheon BBN Technologies,
Cambridge, MA

Raytheon
BBN Technologies
The FTC asked Congress today for additional authority to fight unwanted Internet spam, which now accounts for up to half of all e-mail traffic.

The Federal Trade Commission asked Congress yesterday for broader powers to attack the rapidly growing problem of spam, which new studies show accounts for half of all e-mail traffic.
The FTC asked Congress today for **additional authority** to fight unwanted Internet spam, which now accounts for up to half of all e-mail traffic.

The Federal Trade Commission asked Congress yesterday for **broader powers** to attack the rapidly growing problem of spam, which new studies show accounts for half of all e-mail traffic.
The FTC asked Congress today for **additional authority** to fight unwanted Internet spam, which now accounts for up to half of all e-mail traffic.

The Federal Trade Commission asked Congress yesterday for **broader powers** to attack the rapidly growing problem of spam, which new studies show accounts for half of all e-mail traffic.

Word Similarity
The FTC asked Congress today for **additional authority** to fight unwanted Internet spam, which now accounts for up to half of all e-mail traffic.

The Federal Trade Commission asked Congress yesterday for **broader powers** to attack the rapidly growing problem of spam, which new studies show accounts for half of all e-mail traffic.

---

**Word Similarity**

**Phrasal Similarity**
The FTC asked Congress today for additional authority to fight unwanted Internet spam, which now accounts for up to half of all e-mail traffic.

The Federal Trade Commission asked Congress yesterday for broader powers to attack the rapidly growing problem of spam, which new studies show accounts for half of all e-mail traffic.
● The **FTC** asked Congress **today** for additional authority to fight unwanted Internet spam, which now accounts for up to half of all e-mail traffic.

● The **Federal Trade Commission** asked Congress **yesterday** for broader powers to attack the rapidly growing problem of spam, which new studies show accounts for half of all e-mail traffic.
The FTC asked Congress today for additional authority to fight unwanted Internet spam, which now accounts for up to half of all e-mail traffic.

The Federal Trade Commission asked Congress yesterday for broader powers to attack the rapidly growing problem of spam, which new studies show accounts for half of all e-mail traffic.

**Alias & Date Normalization**
The **FTC** asked Congress **today** for additional authority to fight unwanted Internet spam, which now accounts for up to half of all e-mail traffic.

The **Federal Trade Commission** asked Congress **yesterday** for broader powers to attack the rapidly growing problem of spam, which new studies show accounts for half of all e-mail traffic.

**Alias & Date Normalization**

**Named Entity Equivalence**
The **FTC** asked Congress *today* for additional authority to fight unwanted Internet spam, which now accounts for up to half of all e-mail traffic.

The **Federal Trade Commission** asked Congress *yesterday* for broader powers to attack the rapidly growing problem of spam, which new studies show accounts for half of all e-mail traffic.
They also found shortness was associated with a family history of hearing loss.

Shortness was found twice as often in those with hearing loss.

John Hickenlooper had 65 percent of the vote to 35 percent for City Auditor Don Mares.

Hickenlooper clobbered city Auditor Don Mares, 46, in the Tuesday runoff.
They also found shortness was **associated with** a family history of hearing loss.

Shortness was **found twice as often** in those with hearing loss.

John Hickenlooper had **65 percent of the vote to 35 percent** for City Auditor Don Mares.

Hickenlooper **clobbered** city Auditor Don Mares, 46, in the Tuesday runoff.
They also found shortness was associated with a family history of hearing loss.

Shortness was found twice as often in those with hearing loss.

John Hickenlooper had 65 percent of the vote to 35 percent for City Auditor Don Mares.

Hickenlooper clobbered city Auditor Don Mares, 46, in the Tuesday runoff.

Inference
Which semantic components contribute to STS?

1. **sub-Sentence Similarity**
   - Word Similarity (*distributional, Word Sense, ...*)
   - Phrase/Constituent Similarity (*MWE, ...*)
Which semantic components contribute to STS?

1. **sub-Sentence Similarity**
   - Word Similarity (*distributional, Word Sense, ...*)
   - Phrase/Constituent Similarity (*MWE, ...*)

2. **Equivalent Reorderings**
   - Syntax (*phrase structure parsing, dependency parsing, ...*)
   - Predicate argument structure (*SRL*)

---

Sameer Pradhan
DARPA: Semantic Textual Similarity Workshop
Which semantic components contribute to STS?

1. **sub-Sentence Similarity**
   - Word Similarity (*distributional, Word Sense, ...*)
   - Phrase/Constituent Similarity (*MWE, ...*)

2. **Equivalent Reorderings**
   - Syntax (*phrase structure parsing, dependency parsing, ...*)
   - Predicate argument structure (*SRL*)

3. **Entity/Event Similarity**
   - Coreference (*within-document, cross-document*)
   - World Knowledge

Sameer Pradhan
DARPA: Semantic Textual Similarity Workshop
Which semantic components contribute to STS?

1. **sub-Sentence Similarity**
   - Word Similarity (*distributional, Word Sense, ...*)
   - Phrase/Constituent Similarity (*MWE, ...*)

2. **Equivalent Reorderings**
   - Syntax (*phrase structure parsing, dependency parsing, ...*)
   - Predicate argument structure (*SRL*)

3. **Entity/Event Similarity**
   - Coreference (*within-document, cross-document*)
   - World Knowledge

4. **Focus/Salience**

5. **Negation**

6. **Scope**

7. ...

8. ...