Spoken Content Retrieval: Challenges, Techniques and Applications

(Part 3: Combining IR and ASR)

Gareth J. F. Jones

Centre for Next Generation Localisation
School of Computing, Dublin City University, Dublin, Ireland

Overview

Introduction

Interaction of ASR Error and IR

Word Distributions in accurate vs ASR transcripts

Term weighting and SCR

Exploiting Multiple Hypotheses

Introduction

- The interaction between IR and ASR is non-trivial.

Introduction

- The interaction between IR and ASR is non-trivial.
- ASR errors impact on retrieval behaviour,

Introduction

- The interaction between IR and ASR is non-trivial.
- ASR errors impact on retrieval behaviour, in sometimes surprising ways.

Introduction

- The interaction between IR and ASR is non-trivial.
- ASR errors impact on retrieval behaviour, in sometimes surprising ways.
- Experience has shown that SCR systems can be effective with relatively high WER rates without much apparent effect on performance.

Introduction

- The interaction between IR and ASR is non-trivial.
- ASR errors impact on retrieval behaviour, in sometimes surprising ways.
- Experience has shown that SCR systems can be effective with relatively high WER rates without much apparent effect on performance.
  - Performance largely unaffected with WER of 20%.

Introduction

- The interaction between IR and ASR is non-trivial.
- ASR errors impact on retrieval behaviour, in sometimes surprising ways.
- Experience has shown that SCR systems can be effective with relatively high WER rates without much apparent effect on performance.
  - Performance largely unaffected with WER of 20%.
  - but WER of 30% - 50% or more is likely for many SCR content sets, e.g. call-center recordings, telephone speech, etc.

Introduction

- WER is not a direct predictor of SCR effectiveness,

Introduction

- WER is not a direct predictor of SCR effectiveness, but high WER does impact on SCR quality.

Introduction

- WER is not a direct predictor of SCR effectiveness, but high WER does impact on SCR quality.
  Consider what is going on here:

Introduction

- WER is not a direct predictor of SCR effectiveness, but high WER does impact on SCR quality.
  Consider what is going on here:
  - function words vs meaning-bearing words

Introduction

- WER is not a direct predictor of SCR effectiveness, but high WER does impact on SCR quality. Consider what is going on here:
  - function words vs meaning-bearing words
  - consider only mean-bearing words?

Introduction

- WER is not a direct predictor of SCR effectiveness, but high WER does impact on SCR quality. Consider what is going on here:
  - function words vs meaning-bearing words
  - consider only mean-bearing words? $\times$

Introduction

- WER is not a direct predictor of SCR effectiveness, but high WER does impact on SCR quality.
  Consider what is going on here:
  - function words vs meaning-bearing words
  - consider only mean-bearing words? $\times$
  - consider only named entities?

Introduction

- WER is not a direct predictor of SCR effectiveness, but high WER does impact on SCR quality. Consider what is going on here:
  - function words vs meaning-bearing words
  - consider only mean-bearing words? $\times$
  - consider only named entities? $\checkmark$

Interaction of ASR Error and IR

- Analysis of ranked lists produced in SCR shows that documents at nearer the top of the list have lower WERs than the average for the collection.

Interaction of ASR Error and IR

- Analysis of ranked lists produced in SCR shows that documents at nearer the top of the list have lower WERs than the average for the collection.
- A query expresses a topic. This is better represented if the ASR WER is low and the query can more easily match the contents.

Interaction of ASR Error and IR

- Analysis of ranked lists produced in SCR shows that documents at nearer the top of the list have lower WERs than the average for the collection.
- A query expresses a topic. This is better represented if the ASR WER is low and the query can more easily match the contents.
- Errors in transcriptions are likely to be independent and not associated with topic word patterns.

Interaction of ASR Error and IR

- Analysis of ranked lists produced in SCR shows that documents at nearer the top of the list have lower WERs than the average for the collection.
- A query expresses a topic. This is better represented if the ASR WER is low and the query can more easily match the contents.
- Errors in transcriptions are likely to be independent and not associated with topic word patterns.
- Unlikely that misrecognized words will appear in a pattern that resembles a topic.

Interaction of ASR Error and IR

- Analysis of ranked lists produced in SCR shows that documents at nearer the top of the list have lower WERs than the average for the collection.
- A query expresses a topic. This is better represented if the ASR WER is low and the query can more easily match the contents.
- Errors in transcriptions are likely to be independent and not associated with topic word patterns.
- Unlikely that misrecognized words will appear in a pattern that resembles a topic.
- Hence well recognized relevant items are likely to appear near the top of the list, and non-relevant items in which the words are not spoken are unlikely to be promoted in rank, relative to that of a perfect transcript.

Word Distributions in accurate vs ASR transcripts

- Comparison of human generated and ASR transcripts using an ASR system well matched to the task revealed:

Word Distributions in accurate vs ASR transcripts

- Comparison of human generated and ASR transcripts using an ASR system well matched to the task revealed:
  - ASR transcripts have much smaller vocabulary than manual transcripts.

Word Distributions in accurate vs ASR transcripts

- Comparison of human generated and ASR transcripts using an ASR system well matched to the task revealed:
  - ASR transcripts have much smaller vocabulary than manual transcripts.
  - Count of observations of word which do appear in ASR transcripts (on average) notably higher than in manul transcripts.

Word Distributions in accurate vs ASR transcripts

- Comparison of human generated and ASR transcripts using an ASR system well matched to the task revealed:
  - ASR transcripts have much smaller vocabulary than manual transcripts.
  - Count of observations of word which do appear in ASR transcripts (on average) notably higher than in manul transcripts.
- This arises due to:

Word Distributions in accurate vs ASR transcripts

- Comparison of human generated and ASR transcripts using an ASR system well matched to the task revealed:
  - ASR transcripts have much smaller vocabulary than manual transcripts.
  - Count of observations of word which do appear in ASR transcripts (on average) notably higher than in manul transcripts.
- This arises due to:
  - OOV issues.

Word Distributions in accurate vs ASR transcripts

- Comparison of human generated and ASR transcripts using an ASR system well matched to the task revealed:
  - ASR transcripts have much smaller vocabulary than manual transcripts.
  - Count of observations of word which do appear in ASR transcripts (on average) notably higher than in manul transcripts.
- This arises due to:
  - OOV issues.
  - Problems with acoustic models and language model mean that ASR system shows bias towards use of some in vocabulary words, and bias against using other ones.

Term weighting and SCR

- SCR uses the same term weighting functions as standard text IR, e.g.

Term weighting and SCR

- SCR uses the same term weighting functions as standard text IR, e.g.

$$idf(i) = \log \frac{N}{n(i)} \quad f(tf(i,j)) = \log(tf(i,j) + 1)$$

$$w(i,j) = idf(i) \times f(tf(i,j))$$

where: $idf(i)$ = inverse document frequency of term i
N = no of documents in the current collection
$n(i)$ = no of documents containing term i
$tf(i,j)$ = no of occurrences of term i in document j
$w(i,j)$ = tf.tdf weighting of term i in document j

Term weighting and SCR

- Non-linear functions of this type mean that the first occurrence of a term is the most important.
- Subsequent occurrences have progressively less impact on document rank.

Very basic ranking function:

$$ms(j) = \sum_{i=0}^{I-1} w(i,j)$$

where $ms(j)$ = query-document matching score of document j
I = vocabulary of all search terms

Term weighting and SCR

- Terms that are OOV clearly have $w(i, j) = 0$.

Term weighting and SCR

- Terms that are OOV clearly have $w(i,j) = 0$.
- Terms which are not favoured by the ASR system will have lower values of $n(i)$ than with a perfect transcript,

Term weighting and SCR

- Terms that are OOV clearly have $w(i, j) = 0$.
- Terms which are not favoured by the ASR system will have lower values of $n(i)$ than with a perfect transcript, i.e. they will have higher $idf(i)$ values and appear more important.

Term weighting and SCR

- Terms that are OOV clearly have $w(i,j) = 0$.
- Terms which are not favoured by the ASR system will have lower values of $n(i)$ than with a perfect transcript, i.e. they will have higher $idf(i)$ values and appear more important.
- Terms which are not favoured by the ASR system will have lower values of $tf(i,f)$ than with a perfect transcript,

Term weighting and SCR

- Terms that are OOV clearly have $w(i,j) = 0$.
- Terms which are not favoured by the ASR system will have lower values of $n(i)$ than with a perfect transcript, i.e. they will have higher $idf(i)$ values and appear more important.
- Terms which are not favoured by the ASR system will have lower values of $tf(i,f)$ than with a perfect transcript, i.e. they will have lower values of $f(tf(i,j))$ and appear less significant.

Term weighting and SCR

- Terms that are OOV clearly have $w(i,j) = 0$.
- Terms which are not favoured by the ASR system will have lower values of $n(i)$ than with a perfect transcript, i.e. they will have higher $idf(i)$ values and appear more important.
- Terms which are not favoured by the ASR system will have lower values of $tf(i,f)$ than with a perfect transcript, i.e. they will have lower values of $f(tf(i,j))$ and appear less significant.
- Even for a small document collection $n(i)$ values will 10s or 100s (often much more),

Term weighting and SCR

- Terms that are OOV clearly have $w(i, j) = 0$.
- Terms which are not favoured by the ASR system will have lower values of $n(i)$ than with a perfect transcript, i.e. they will have higher $idf(i)$ values and appear more important.
- Terms which are not favoured by the ASR system will have lower values of $tf(i, f)$ than with a perfect transcript, i.e. they will have lower values of $f(tf(i, j)))$ and appear less significant.
- Even for a small document collection $n(i)$ values will 10s or 100s (often much more), impact on $idf(i)$ will usually be very small.

Term weighting and SCR

- Terms that are OOV clearly have $w(i,j) = 0$.
- Terms which are not favoured by the ASR system will have lower values of $n(i)$ than with a perfect transcript, i.e. they will have higher $idf(i)$ values and appear more important.
- Terms which are not favoured by the ASR system will have lower values of $tf(i,f)$ than with a perfect transcript, i.e. they will have lower values of $f(tf(i,j))$ and appear less significant.
- Even for a small document collection $n(i)$ values will 10s or 100s (often much more), impact on $idf(i)$ will usually be very small.
- Impact on $ms(i,j)$ will be minimal.

Term weighting and SCR

- The **real** $tf_R(i, j)$ for a word relevant to document j will typically be > 1.

Term weighting and SCR

- The **real** $tf_R(i,j)$ for a word relevant to document j will typically be > 1.
- If ASR produces $tf(i,j) = 0$, when $tf_R(i,j) > 0$, then clearly the mismatch is a problem.

Term weighting and SCR

- The **real** $tf_R(i,j)$ for a word relevant to document j will typically be > 1.
- If ASR produces $tf(i,j) = 0$, when $tf_R(i,j) > 0$, then clearly the mismatch is a problem.
- If ASR produces $tf(i,j) < tf_R(i,j)$, then $w(i,j)$ and $ms(j)$ will be reduced,

Term weighting and SCR

- The **real** $tf_R(i,j)$ for a word relevant to document j will typically be > 1.
- If ASR produces $tf(i,j) = 0$, when $tf_R(i,j) > 0$, then clearly the mismatch is a problem.
- If ASR produces $tf(i,j) < tf_R(i,j)$, then $w(i,j)$ and $ms(j)$ will be reduced, but not usually not disasterously,

Term weighting and SCR

- The **real** $tf_R(i,j)$ for a word relevant to document j will typically be > 1.
- If ASR produces $tf(i,j) = 0$, when $tf_R(i,j) > 0$, then clearly the mismatch is a problem.
- If ASR produces $tf(i,j) < tf_R(i,j)$, then $w(i,j)$ and $ms(j)$ will be reduced, but not usually not disasterously, i.e SCR is robust to some level of substitutions and deletions of correct terms.

Term weighting and SCR

- The **real** $tf_R(i,j)$ for a word relevant to document j will typically be > 1.
- If ASR produces $tf(i,j) = 0$, when $tf_R(i,j) > 0$, then clearly the mismatch is a problem.
- If ASR produces $tf(i,j) < tf_R(i,j)$, then $w(i,j)$ and $ms(j)$ will be reduced, but not usually not disasterously, i.e SCR is robust to some level of substitutions and deletions of correct terms.
- By contrast, insertions and substitutions of incorrect term will typically have $tf(i,j) = 1$ in j,
  AND other terms in the query will typically have $tf(i,j) = 0$.
  Thus, $ms(i,j)$ for these documents will be > 0, BUT will typically be very low, and may fall below a document score threshold to be retrieved.

Term weighting and SCR

- Words which are favoured by the ASR system often have $n(i) >> n_R(i)$, where $n_R(i)$ is the **real** no of documents containing term i.

Term weighting and SCR

- Words which are favoured by the ASR system often have $n(i) >> n_R(i)$, where $n_R(i)$ is the **real** no of documents containing term i.
  $idf(i)$ will be reduced,

Term weighting and SCR

- Words which are favoured by the ASR system often have $n(i) >> n_R(i)$, where $n_R(i)$ is the **real** no of documents containing term i.
  $idf(i)$ will be reduced, but generally not significantly.

Term weighting and SCR

- Words which are favoured by the ASR system often have $n(i) >> n_R(i)$, where $n_R(i)$ is the **real** no of documents containing term i.
  $idf(i)$ will be reduced, but generally not significantly.
  $ms(i,j)$ will be reduced, but generally not significantly.
  i.e. term weighting in SCR is quite robust to high levels of insertions and substitutions favouring a term.

Term weighting and SCR

- Words which are favoured by the ASR system often have $n(i) >> n_R(i)$, where $n_R(i)$ is the **real** no of documents containing term i.
  $idf(i)$ will be reduced, but generally not significantly.
  $ms(i, j)$ will be reduced, but generally not significantly.
  i.e. term weighting in SCR is quite robust to high levels of insertions and substitutions favouring a term.
- Words which are favoured by ASR system may have increased $tf(i, j)$, for term i in document j,

Term weighting and SCR

- Words which are favoured by the ASR system often have $n(i) >> n_R(i)$, where $n_R(i)$ is the **real** no of documents containing term i.
  $idf(i)$ will be reduced, but generally not significantly.
  $ms(i, j)$ will be reduced, but generally not significantly.
  i.e. term weighting in SCR is quite robust to high levels of insertions and substitutions favouring a term.
- Words which are favoured by ASR system may have increased $tf(i, j)$, for term i in document j, but since errors are (in a loose sense) randomly distributed this is unlikely.

Term weighting and SCR

- Words which are favoured by the ASR system often have $n(i) >> n_R(i)$, where $n_R(i)$ is the **real** no of documents containing term i.
  $idf(i)$ will be reduced, but generally not significantly.
  $ms(i,j)$ will be reduced, but generally not significantly.
  i.e. term weighting in SCR is quite robust to high levels of insertions and substitutions favouring a term.
- Words which are favoured by ASR system may have increased $tf(i,j)$, for term i in document j, but since errors are (in a loose sense) randomly distributed this is unlikely.
- If $tf(i,f) > tf_R(i,j)$, $f(tf(i,j))$ will be increased, but not significantly, remember the function is usually non-linear.

Term weighting and SCR

- Words which are favoured by the ASR system often have $n(i) >> n_R(i)$, where $n_R(i)$ is the **real** no of documents containing term i.
  $idf(i)$ will be reduced, but generally not significantly.
  $ms(i,j)$ will be reduced, but generally not significantly.
  i.e. term weighting in SCR is quite robust to high levels of insertions and substitutions favouring a term.
- Words which are favoured by ASR system may have increased $tf(i,j)$, for term i in document j, but since errors are (in a loose sense) randomly distributed this is unlikely.
- If $tf(i,f) > tf_R(i,j)$, $f(tf(i,j))$ will be increased, but not significantly, remember the function is usually non-linear.
  $ms(i,j)$ will be increased, but not significantly.

Substituting *idf*(*n*)

- Text document collections are typically much larger than spoken content collections.

Substituting *idf*(n)

- Text document collections are typically much larger than spoken content collections.
- $n(i)$ values for large collections will typically be more representative of a language than for small ones.

Substituting $idf(n)$

- Text document collections are typically much larger than spoken content collections.
- $n(i)$ values for large collections will typically be more representative of a language than for small ones.
- SCR effectiveness may be improved by substituting $idf(i)$ values for a large text collection for the $idf(i)$ values calculated using ASR transcripts,

Substituting *idf*(*n*)

- Text document collections are typically much larger than spoken content collections.
- $n(i)$ values for large collections will typically be more representative of a language than for small ones.
- SCR effectiveness may be improved by substituting $idf(i)$ values for a large text collection for the $idf(i)$ values calculated using ASR transcripts,

  BUT the text collection **MUST** be strongly topically or temporally related to the SCR collection.

Substituting *idf*(n)

- Text document collections are typically much larger than spoken content collections.
- $n(i)$ values for large collections will typically be more representative of a language than for small ones.
- SCR effectiveness may be improved by substituting $idf(i)$ values for a large text collection for the $idf(i)$ values calculated using ASR transcripts,

  BUT the text collection **MUST** be strongly topically or temporally related to the SCR collection.

  - Has been demonstrated to be effective for SCR for news data in TREC SDR tasks.

Substituting *idf*(n)

- Text document collections are typically much larger than spoken content collections.
- $n(i)$ values for large collections will typically be more representative of a language than for small ones.
- SCR effectiveness may be improved by substituting $idf(i)$ values for a large text collection for the $idf(i)$ values calculated using ASR transcripts,

  BUT the text collection **MUST** be strongly topically or temporally related to the SCR collection.

  - Has been demonstrated to be effective for SCR for news data in TREC SDR tasks.
  - BUT, text data must be contemporaneous with the spoken data

Substituting $idf(n)$

- Text document collections are typically much larger than spoken content collections.
- $n(i)$ values for large collections will typically be more representative of a language than for small ones.
- SCR effectiveness may be improved by substituting $idf(i)$ values for a large text collection for the $idf(i)$ values calculated using ASR transcripts,

  BUT the text collection **MUST** be strongly topically or temporally related to the SCR collection.

  - Has been demonstrated to be effective for SCR for news data in TREC SDR tasks.
  - BUT, text data must be contemporaneous with the spoken data - same people, places, news events, etc.

Substituting $idf(n)$

- The same $idf(i)$ substitution has been demonstrated to be effective for small text collections - see TREC ad hoc retrieval tasks.

$n_R(i)$ and $tf_R(i, j)$ for text documents

- By contrast consider the situation for **real** values in actual natural text documents.

$n_R(i)$ and $tf_R(i,j)$ for text documents

- By contrast consider the situation for **real** values in actual natural text documents.
- Many documents contain typographical errors - even professionally edited documents such as published news articles.

$n_R(i)$ and $tf_R(i,j)$ for text documents

- By contrast consider the situation for **real** values in actual natural text documents.
- Many documents contain typographical errors - even professionally edited documents such as published news articles.
- Likely to introduce **new** terms into vocabulary.

$n_R(i)$ and $tf_R(i, j)$ for text documents

- By contrast consider the situation for **real** values in actual natural text documents.
- Many documents contain typographical errors - even professionally edited documents such as published news articles.
- Likely to introduce **new** terms into vocabulary.
  - May be filtered out as too rare to include in document index.

$n_R(i)$ and $tf_R(i, j)$ for text documents

- By contrast consider the situation for **real** values in actual natural text documents.
- Many documents contain typographical errors - even professionally edited documents such as published news articles.
- Likely to introduce **new** terms into vocabulary.
  - May be filtered out as too rare to include in document index.
  - Otherwise will have very high $idf(i)$ value, but often lower $tf(i, j)$ value.

$n_R(i)$ and $tf_R(i,j)$ for text documents

- By contrast consider the situation for **real** values in actual natural text documents.
- Many documents contain typographical errors - even professionally edited documents such as published news articles.
- Likely to introduce **new** terms into vocabulary.
  - May be filtered out as too rare to include in document index.
  - Otherwise will have very high $idf(i)$ value, but often lower $tf(i,j)$ value.
  - Thus, high $w(i,j)$ value.

$n_R(i)$ and $tf_R(i,j)$ for text documents

- By contrast consider the situation for **real** values in actual natural text documents.
- Many documents contain typographical errors - even professionally edited documents such as published news articles.
- Likely to introduce **new** terms into vocabulary.
  - May be filtered out as too rare to include in document index.
  - Otherwise will have very high $idf(i)$ value, but often lower $tf(i,j)$ value.
  - Thus, high $w(i,j)$ value.
  - Since they are a typo, usually doesn't matter, they do appear in queries, unless the searcher makes the same typo!

Motivating Use of Multiple Hypotheses

- User queries are typically short - 2 or 3 words.

Motivating Use of Multiple Hypotheses

- User queries are typically short - 2 or 3 words.
- ASR transcripts have deletions and subsitutions.

Motivating Use of Multiple Hypotheses

- User queries are typically short - 2 or 3 words.
- ASR transcripts have deletions and subsitutions.
- $ms(i, j)$ query-document matching score may be very poor where one or more of the query words is poorly recognized or OOV.

Motivating Use of Multiple Hypotheses

- User queries are typically short - 2 or 3 words.
- ASR transcripts have deletions and subsitutions.
- $ms(i, j)$ query-document matching score may be very poor where one or more of the query words is poorly recognized or OOV.
- $ms(i, j)$ should be more reliable, if we can add missing terms to the transcript.

Motivating Use of Multiple Hypotheses

- User queries are typically short - 2 or 3 words.
- ASR transcripts have deletions and subsitutions.
- $ms(i,j)$ query-document matching score may be very poor where one or more of the query words is poorly recognized or OOV.
- $ms(i,j)$ should be more reliable, if we can add missing terms to the transcript.
- Exploiting multiple hypotheses is a potential way to address missing in vocabulary terms,

Motivating Use of Multiple Hypotheses

- User queries are typically short - 2 or 3 words.
- ASR transcripts have deletions and subsitutions.
- $ms(i,j)$ query-document matching score may be very poor where one or more of the query words is poorly recognized or OOV.
- $ms(i,j)$ should be more reliable, if we can add missing terms to the transcript.
- Exploiting multiple hypotheses is a potential way to address missing in vocabulary terms,

  but we need different method to address OOV.

Exploiting Multiple Hypotheses

- N-best lists, Word Lattices and Confusion Networks - have the potential to include correct words substituted out or deleted from the 1-best list,

Exploiting Multiple Hypotheses

- N-best lists, Word Lattices and Confusion Networks - have the potential to include correct words substituted out or deleted from the 1-best list, but also, inserting incorrect words.

Exploiting Multiple Hypotheses

- N-best lists, Word Lattices and Confusion Networks - have the potential to include correct words substituted out or deleted from the 1-best list, but also, inserting incorrect words.
- To improve the situation with respect to substitutions and deletions, the depth of the N-best list, Word Lattice or Confusion Network.

Exploiting Multiple Hypotheses

- N-best lists, Word Lattices and Confusion Networks - have the potential to include correct words substituted out or deleted from the 1-best list, but also, inserting incorrect words.
- To improve the situation with respect to substitutions and deletions, the depth of the N-best list, Word Lattice or Confusion Network.
- But remember, ASR systems are biased against some words and in favour of others.

Exploiting Multiple Hypotheses

- N-best lists, Word Lattices and Confusion Networks - have the potential to include correct words substituted out or deleted from the 1-best list, but also, inserting incorrect words.
- To improve the situation with respect to substitutions and deletions, the depth of the N-best list, Word Lattice or Confusion Network.
- But remember, ASR systems are biased against some words and in favour of others.
- Thus, in order to overcome substitutions and deletions a depth may need to be very deep,

Exploiting Multiple Hypotheses

- N-best lists, Word Lattices and Confusion Networks - have the potential to include correct words substituted out or deleted from the 1-best list, but also, inserting incorrect words.
- To improve the situation with respect to substitutions and deletions, the depth of the N-best list, Word Lattice or Confusion Network.
- But remember, ASR systems are biased against some words and in favour of others.
- Thus, in order to overcome substitutions and deletions a depth may need to be very deep,
- but this risks introducing very high levels of insertions of favoured words.

Exploiting Multiple Hypotheses

- N-best lists, Word Lattices and Confusion Networks - have the potential to include correct words substituted out or deleted from the 1-best list, but also, inserting incorrect words.
- To improve the situation with respect to substitutions and deletions, the depth of the N-best list, Word Lattice or Confusion Network.
- But remember, ASR systems are biased against some words and in favour of others.
- Thus, in order to overcome substitutions and deletions a depth may need to be very deep,
- but this risks introducing very high levels of insertions of favoured words.
- Depth of N-best, Word Lattice or Confusion Network thus represent a trade off.

Exploiting Multiple Hypotheses

- Consider the implications for term weighting.

Exploiting Multiple Hypotheses

- Consider the implications for term weighting.
- $n(i)$ values - for both correctly and incorrectly hypothesized words - will be increased.

  $idf(i)$ values will be decreased.

Exploiting Multiple Hypotheses

- Consider the implications for term weighting.
- $n(i)$ values - for both correctly and incorrectly hypothesized words - will be increased.

  $idf(i)$ values will be decreased.
- $tf(i,j)$ values will increase.
  - As depth increased likelihood of $tf(i,j) > 0$, when $tf_R(i,j) = 0$ will also increase.
  - Due to ASR system bias, this increase may be non-linear with increase in depth.
  - Thus, depth needs to be carefully determined.

Word Confidence Measures

- Since ASR is statistical, we can calculate a likelihood of correct recognition for individual words.
  - Various methods have been proposed to do this.

Word Confidence Measures

- Since ASR is statistical, we can calculate a likelihood of correct recognition for individual words.
  - Various methods have been proposed to do this.
- How might we use this confidence information to improve SCR?

Word Confidence Measures

- Since ASR is statistical, we can calculate a likelihood of correct recognition for individual words.
  - Various methods have been proposed to do this.
- How might we use this confidence information to improve SCR?
- Two approaches:

Word Confidence Measures

- Since ASR is statistical, we can calculate a likelihood of correct recognition for individual words.
  - Various methods have been proposed to do this.
- How might we use this confidence information to improve SCR?
- Two approaches:
  - Thresholding

Word Confidence Measures

- Since ASR is statistical, we can calculate a likelihood of correct recognition for individual words.
  - Various methods have been proposed to do this.
- How might we use this confidence information to improve SCR?
- Two approaches:
  - Thresholding
  - Incorporate in term weights

Word Confidence Measures

Thresholding:

- Compute confidence score for each word hypothesis.

Word Confidence Measures

Thresholding:

- Compute confidence score for each word hypothesis.
  - Since can be different from ASR transcripts score - can overcome biasing effect.

Word Confidence Measures

Thresholding:

- Compute confidence score for each word hypothesis.
  - Since can be different from ASR transcripts score - can overcome biasing effect.
  - Delete words which fall below threshold.

Word Confidence Measures

Thresholding:

- Compute confidence score for each word hypothesis.
  - Since can be different from ASR transcripts score - can overcome biasing effect.
  - Delete words which fall below threshold.
  - Again trade off between insertions, substitutions and deletions.

Word Confidence Measures

Thresholding:

- Compute confidence score for each word hypothesis.
  - Since can be different from ASR transcripts score - can overcome biasing effect.
  - Delete words which fall below threshold.
  - Again trade off between insertions, substitutions and deletions.
  - Typically only use for N-best lists or Word Lattices - not really a problem for 1-best.

Word Confidence Measures

Thresholding:

- Compute confidence score for each word hypothesis.
  - Since can be different from ASR transcripts score - can overcome biasing effect.
  - Delete words which fall below threshold.
  - Again trade off between insertions, substitutions and deletions.
  - Typically only use for N-best lists or Word Lattices - not really a problem for 1-best.
- Alternative for N-best lists is simply to sum them, since insertion issues is not significant.

Word Confidence Measures

Combining with Term Weights:

$$idf(i) = \log \frac{N}{n(i)} \quad f(tf(i,j)) = \log(tf(i,j) + 1)$$

$$w(i,j) = idf(i) \times f(tf(i,j))$$

Word Confidence Measures

Combining with Term Weights:

$$idf(i) = \log \frac{N}{n(i)} \quad f(tf(i,j)) = \log(tf(i,j) + 1)$$

$$w(i,j) = idf(i) \times f(tf(i,j))$$

- Replace $n(i)$ and/or $tf(i,j)$ with a confidence based measure, e.g.
  - Sum of confidence for each word.
  - Estimated count taking account of recognition behaviour.

Out-of-Vocabulary (OOV) Issues

- Use subword recognition method to construct mid-level output at recognition stage.

Out-of-Vocabulary (OOV) Issues

- Use subword recognition method to construct mid-level output at recognition stage.
- Look for OOV query words in subword structure.

Out-of-Vocabulary (OOV) Issues

- Use subword recognition method to construct mid-level output at recognition stage.
- Look for OOV query words in subword structure.
- Methods:

Out-of-Vocabulary (OOV) Issues

- Use subword recognition method to construct mid-level output at recognition stage.
- Look for OOV query words in subword structure.
- Methods:
  - Phone lattice spotting

Out-of-Vocabulary (OOV) Issues

- Use subword recognition method to construct mid-level output at recognition stage.
- Look for OOV query words in subword structure.
- Methods:
  - Phone lattice spotting
    - scalability problems

Out-of-Vocabulary (OOV) Issues

- Use subword recognition method to construct mid-level output at recognition stage.
- Look for OOV query words in subword structure.
- Methods:
  - Phone lattice spotting
    - scalability problems
  - Spoken Term Detection

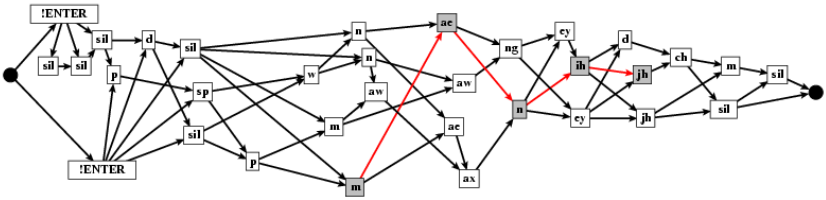
Out-of-Vocabulary (OOV) Issues

- Use subword recognition method to construct mid-level output at recognition stage.
- Look for OOV query words in subword structure.
- Methods:
  - Phone lattice spotting
    - scalability problems
  - Spoken Term Detection
    - Designed to be accurate and scalable

Out-of-Vocabulary (OOV) Issues

- Use subword recognition method to construct mid-level output at recognition stage.
- Look for OOV query words in subword structure.
- Methods:
  - Phone lattice spotting
    - scalability problems
  - Spoken Term Detection
    - Designed to be accurate and scalable
- Alternative strategy, 1-best ASR transcript, but look for ALL query terms using OOV system to try to overcome deletion and substitutions issues, rather than N-best list or word lattice.

Phone Lattice Spotting



- Again a trade off in depth, confidence measures typically used to filter hypotheses.

Combining Index Sources

Two basic alternative:

Combining Index Sources

Two basic alternative:

- ASR transcripts and output of OOV recognition can be combined into a single index for a document.

Combining Index Sources

Two basic alternative:

- ASR transcripts and output of OOV recognition can be combined into a single index for a document.
  - Consider effects on term weights.

Combining Index Sources

Two basic alternative:

- ASR transcripts and output of OOV recognition can be combined into a single index for a document.
  - Consider effects on term weights.
- ASR transcripts can be used to construct separate index files for different sources.

Combining Index Sources

Two basic alternative:

- ASR transcripts and output of OOV recognition can be combined into a single index for a document.
  - Consider effects on term weights.
- ASR transcripts can be used to construct separate index files for different sources.

  Compute separate $ms(j)$ for each index.

Combining Index Sources

Two basic alternative:

- ASR transcripts and output of OOV recognition can be combined into a single index for a document.
  - Consider effects on term weights.
- ASR transcripts can be used to construct separate index files for different sources.

  Compute separate $ms(j)$ for each index.

  Form weighted linear sum of scores.

Combining Index Sources

Two basic alternative:

- ASR transcripts and output of OOV recognition can be combined into a single index for a document.
  - Consider effects on term weights.
- ASR transcripts can be used to construct separate index files for different sources.

  Compute separate $ms(j)$ for each index.

  Form weighted linear sum of scores.

  Form final ranked list.

Combining Index Sources

Two basic alternative:

- ASR transcripts and output of OOV recognition can be combined into a single index for a document.
  - Consider effects on term weights.
- ASR transcripts can be used to construct separate index files for different sources.

  Compute separate $ms(j)$ for each index.

  Form weighted linear sum of scores.

  Form final ranked list.