Language Technology: Research and Development

Dissemination of Research Results

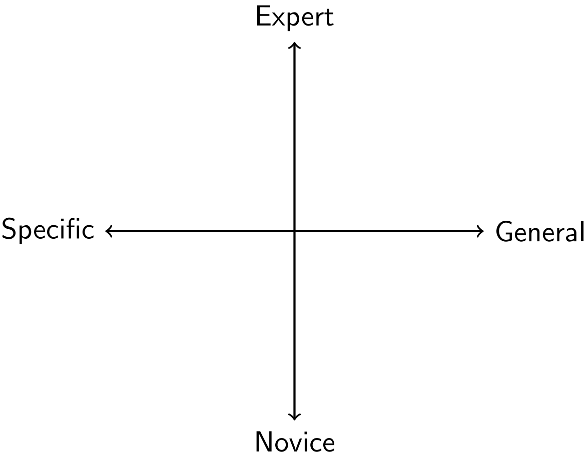
Sara Stymne

Uppsala University
Department of Linguistics and Philology
sara.stymne@lingfil.uu.se

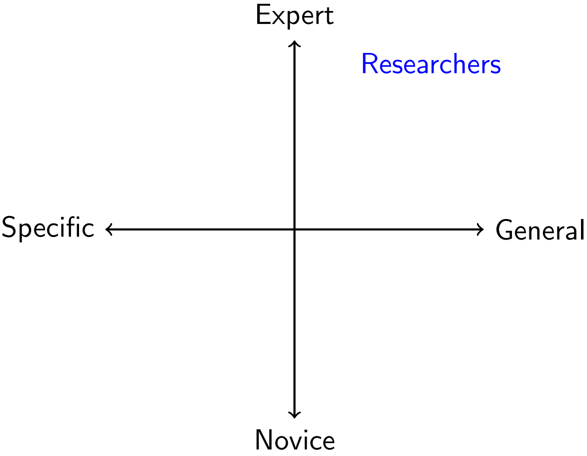
Dissemination of Research Results

- Why?
  - Submit results for critical review
  - Inform other researchers, users, society
  - Satisfy requirements from funders or customers
  - Promote research career – publish or perish
- To whom?
  - Other researchers
  - Potential users
  - Students
  - The general public
  - Funding bodies
  - Customers

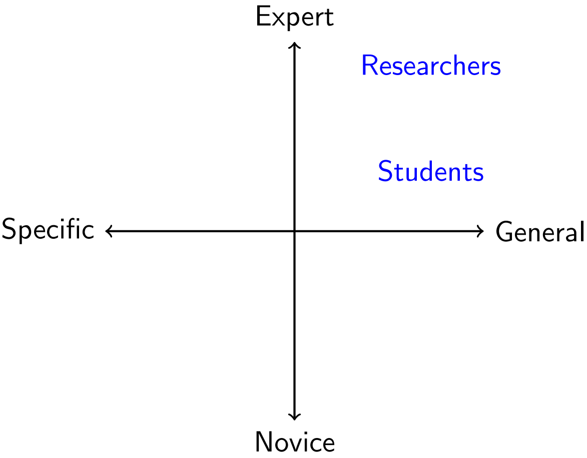
The Receiver



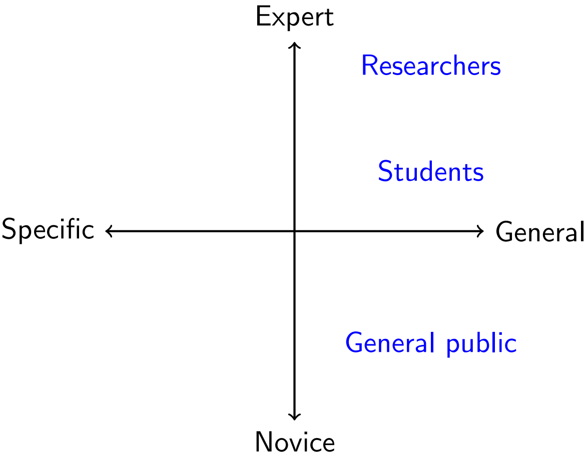
The Receiver



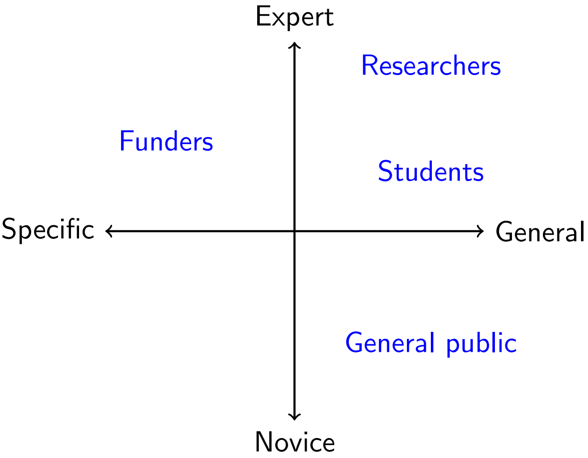
The Receiver



The Receiver



The Receiver



Expert

Researchers

Funders

Students

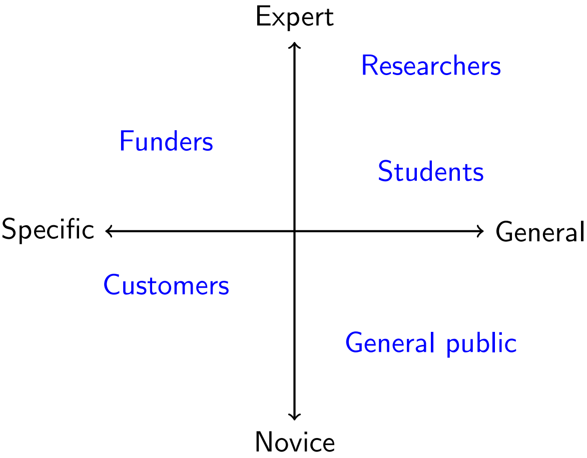
Specific

General

General public

Novice

The Receiver



Expert

Researchers

Funders

Students

Specific

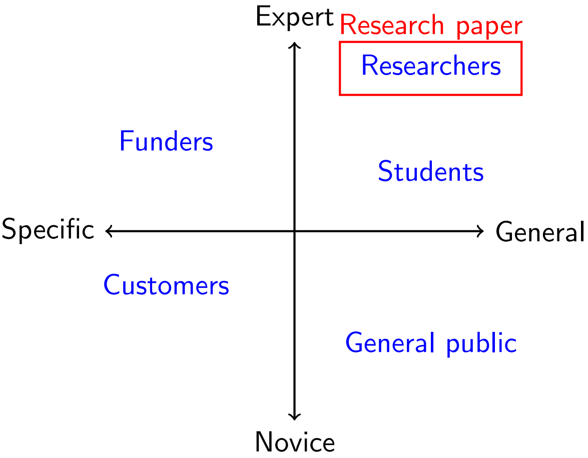
General

Customers

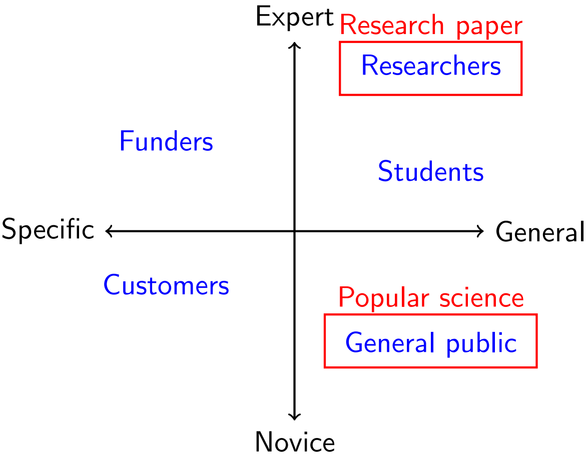
General public

Novice

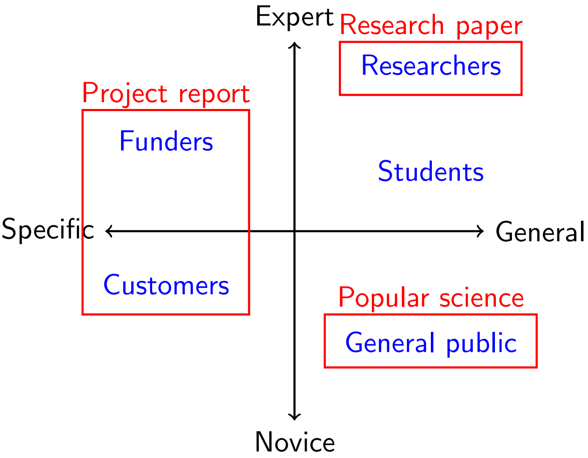
The Receiver



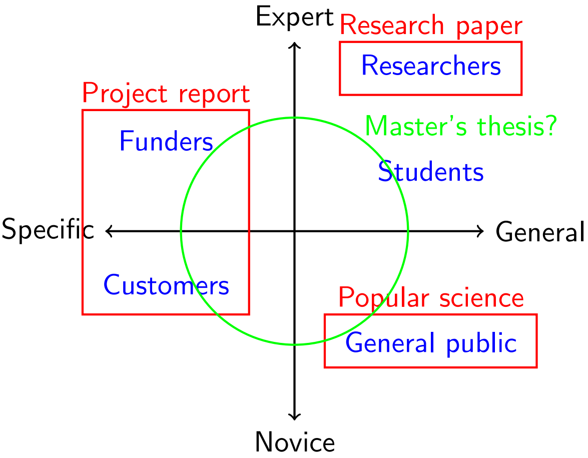
The Receiver



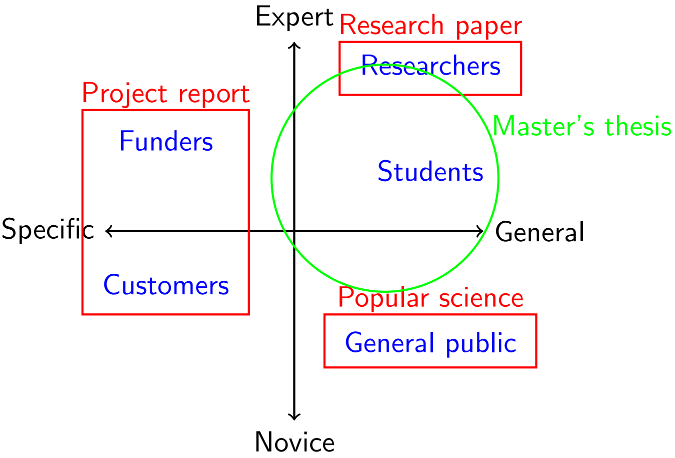
The Receiver



The Receiver



The Receiver



How?

Written:

1. Publications (indexed and archived)
2. Internal reports (public or confidential)
3. Digital archives, web pages, etc.

Oral:

1. Lectures (especially at conferences)
2. Demonstrations, posters, discussions, etc.
3. Internal meetings (seminars, workshops)

Written Genres – Single Topic

Papers (short)

1. Journal article – refereed and approved by editorial board
2. Conference paper – often but not always refereed
3. Technical report – usually not refereed

Monographs (long)

1. Book – standards of refereeing depends on publisher
2. Thesis – refereed in examination, may or may not be published

Written Genres – Other

Collections

1. Conference proceedings – collection of conference papers
2. Edited volume – book with different chapter authors

Meta-genres

1. Survey or handbook article
2. Review in scientific journal
3. Bibliography
4. Abstract

Oral Genres

Lecture

- Presentation by 1 person followed by discussion (large group)
    1. Conference talk (15–30 min)
    2. Invited talk (45–90 min)

Seminar

- Presentation or introduction by 1 or more persons with more or less continous discussion (small group)

Panel

- Short presentations on a set topic from a selected group of persons with questions and opinions from the audience

Mixed Genres

Poster

- Written presentation displayed on poster board
- Oral interaction with interested audience
- Sometimes combined with short talk (1–5 min)

Demonstration

- System demonstration (or similar)
- Oral interaction with interested audience
- Sometimes combined with poster

Requirements on Scientific Reports

- Ethics:
  - Sensitive information requires permission and anonymization
- Accessibility:
  - Reports should be understandable by target audience
- Novelty and relevance:
  - Results should be novel, original, unpublished
  - Relevance to research area should be made clear
- Quality:
  - Claims clearly stated and possible to challenge (falsifiability)
  - Claims supported by arguments and/or evidence (justification)
  - Claims not misleading (e.g., by withholding information)

Scientific Writing

Writing takes time (to learn)

- Practice makes perfect – write a lot!
- Writing requires rewriting – start early!

Scientific writing is a standardized genre

- Collect good examples – and study them!
- Copy structure and formulations – but not content!

The Structure of Scientific Publications

The Structure of Scientific Publications

Pre-matter: Title page (abstract, preface, contents)

Post-matter: References (appendices, indexes)

The Structure of Scientific Publications

Pre-matter: Title page (abstract, preface, contents)

Introduction: What is the problem/question?
Why is it relevant/interesting?

Conclusion: What is the solution/answer?
Where do we go from here?

Post-matter: References (appendices, indexes)

The Structure of Scientific Publications

Pre-matter: Title page (abstract, preface, contents)

Introduction: What is the problem/question?
Why is it relevant/interesting?

Body: What has been done before?
How is the problem tackled?
What are the results?

Conclusion: What is the solution/answer?
Where do we go from here?

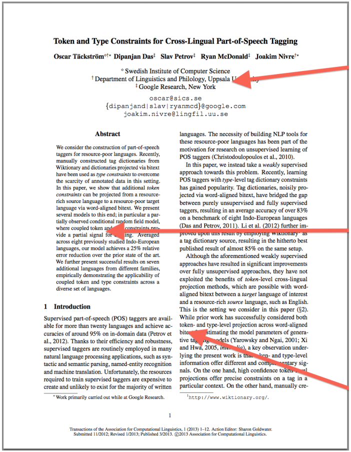
Post-matter: References (appendices, indexes)

The Main Theme

The research question

- is stated in the introduction
- is related to previous research
- motivates the approach taken
- determines the selection of results
- is revisited in the conclusion

The Anatomy of a TACL Style Article



Token and Type Constraints for Cross-Lingual Part-of-Speech Tagging

Oscar Täckström[◇†*] **Dipanjan Das**[‡] **Slav Petrov**[‡] **Ryan McDonald**[‡] **Joakim Nivre**[†*]

[◇] Swedish Institute of Computer Science
[†] Department of Linguistics and Philology, Uppsala University
[‡] Google Research, New York

oscar@sics.se
{dipanjand|slav|ryanmcd}@google.com
joakim.nivre@lingfil.uu.se

Abstract

We consider the construction of part-of-speech taggers for resource-poor languages. Recently, manually constructed tag dictionaries from Wiktionary and dictionaries projected via bitext have been used as *type constraints* to overcome the scarcity of annotated data in this setting. In this paper, we show that additional *token constraints* can be projected from a resource-rich source language to a resource-poor target language via word-aligned bitext. We present several models to this end; in particular a partially observed conditional random field model, where coupled token and type constraints provide a partial signal for training. Averaged across eight previously studied Indo-European languages, our model achieves a 25% relative error reduction over the prior state of the art. We further present successful results on seven additional languages from different families, empirically demonstrating the applicability of coupled token and type constraints across a diverse set of languages.

1 Introduction

Supervised part-of-speech (POS) taggers are available for more than twenty languages and achieve accuracies of around 95% on in-domain data (Petrov et al., 2012). Thanks to their efficiency and robustness, supervised taggers are routinely employed in many natural language processing applications, such as syntactic and semantic parsing, named-entity recognition and machine translation. Unfortunately, the resources required to train supervised taggers are expensive to create and unlikely to exist for the majority of written languages. The necessity of building NLP tools for these resource-poor languages has been part of the motivation for research on unsupervised learning of POS taggers (Christodoulopoulos et al., 2010).

In this paper, we instead take a *weakly* supervised approach towards this problem. Recently, learning POS taggers with *type*-level tag dictionary constraints has gained popularity. Tag dictionaries, noisily projected via word-aligned bitext, have bridged the gap between purely unsupervised and fully supervised taggers, resulting in an average accuracy of over 83% on a benchmark of eight Indo-European languages (Das and Petrov, 2011). Li et al. (2012) further improved upon this result by employing Wiktionary[1] as a tag dictionary source, resulting in the hitherto best published result of almost 85% on the same setup.

Although the aforementioned weakly supervised approaches have resulted in significant improvements over fully unsupervised approaches, they have not exploited the benefits of *token*-level cross-lingual projection methods, which are possible with word-aligned bitext between a *target* language of interest and a resource-rich *source* language, such as English. This is the setting we consider in this paper (§2). While prior work has successfully considered both token- and type-level projection across word-aligned bitext for estimating the model parameters of generative tagging models (Yarowsky and Ngai, 2001; Xi and Hwa, 2005, *inter alia*), a key observation underlying the present work is that token- and type-level information offer different and complementary signals. On the one hand, high confidence token-level projections offer precise constraints on a tag in a particular context. On the other hand, manually cre-

[*] Work primarily carried out while at Google Research.

[1] http://www.wiktionary.org/.

1

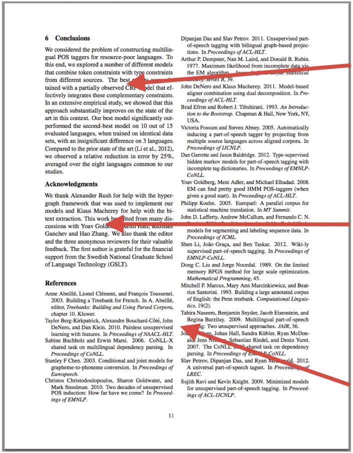
Transactions of the Association for Computational Linguistics, 1 (2013) 1–12. Action Editor: Sharon Goldwater.
Submitted 11/2012; Revised 1/2013; Published 3/2013. ©2013 Association for Computational Linguistics.

Title page: title, authors, affiliations

Abstract: self-contained summary

Main text in numbered sections

The Anatomy of a TACL Style Article



6 Conclusions

We considered the problem of constructing multilingual POS taggers for resource-poor languages. To this end, we explored a number of different models that combine token constraints with type constraints from different sources. The best results were obtained with a partially observed CRF model that effectively integrates these complementary constraints. In an extensive empirical study, we showed that this approach substantially improves on the state of the art in this context. Our best model significantly outperformed the second-best model on 10 out of 15 evaluated languages, when trained on identical data sets, with an insignificant difference on 3 languages. Compared to the prior state of the art (Li et al., 2012), we observed a relative reduction in error by 25%, averaged over the eight languages common to our studies.

Acknowledgments

We thank Alexander Rush for help with the hypergraph framework that was used to implement our models and Klaus Macherey for help with the bitext extraction. This work benefited from many discussions with Yoav Goldberg, Keith Hall, Kuzman Ganchev and Hao Zhang. We also thank the editor and the three anonymous reviewers for their valuable feedback. The first author is grateful for the financial support from the Swedish National Graduate School of Language Technology (GSLT).

References

Anne Abeillé, Lionel Clément, and François Toussenel. 2003. Building a Treebank for French. In A. Abeillé, editor, *Treebanks: Building and Using Parsed Corpora*, chapter 10. Kluwer.

Taylor Berg-Kirkpatrick, Alexandre Bouchard-Côté, John DeNero, and Dan Klein. 2010. Painless unsupervised learning with features. In *Proceedings of NAACL-HLT*.

Sabine Buchholz and Erwin Marsi. 2006. CoNLL-X shared task on multilingual dependency parsing. In *Proceedings of CoNLL*.

Stanley F Chen. 2003. Conditional and joint models for grapheme-to-phoneme conversion. In *Proceedings of Eurospeech*.

Christos Christodoulopoulos, Sharon Goldwater, and Mark Steedman. 2010. Two decades of unsupervised POS induction: How far have we come? In *Proceedings of EMNLP*.

Dipanjan Das and Slav Petrov. 2011. Unsupervised part-of-speech tagging with bilingual graph-based projections. In *Proceedings of ACL-HLT*.

Arthur P. Dempster, Nan M. Laird, and Donald B. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39.

John DeNero and Klaus Macherey. 2011. Model-based aligner combination using dual decomposition. In *Proceedings of ACL-HLT*.

Brad Efron and Robert J. Tibshirani. 1993. *An Introduction to the Bootstrap*. Chapman & Hall, New York, NY, USA.

Victoria Fossum and Steven Abney. 2005. Automatically inducing a part-of-speech tagger by projecting from multiple source languages across aligned corpora. In *Proceedings of IJCNLP*.

Dan Garrette and Jason Baldridge. 2012. Type-supervised hidden markov models for part-of-speech tagging with incomplete tag dictionaries. In *Proceedings of EMNLP-CoNLL*.

Yoav Goldberg, Meni Adler, and Michael Elhadad. 2008. EM can find pretty good HMM POS-taggers (when given a good start). In *Proceedings of ACL-HLT*.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT Summit*.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of ICML*.

Shen Li, João Graça, and Ben Taskar. 2012. Wiki-ly supervised part-of-speech tagging. In *Proceedings of EMNLP-CoNLL*.

Dong C. Liu and Jorge Nocedal. 1989. On the limited memory BFGS method for large scale optimization. *Mathematical Programming*, 45.

Mitchell P. Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. Building a large annotated corpus of English: the Penn treebank. *Computational Linguistics*, 19(2).

Tahira Naseem, Benjamin Snyder, Jacob Eisenstein, and Regina Barzilay. 2009. Multilingual part-of-speech tagging: Two unsupervised approaches. *JAIR*, 36.

Joakim Nivre, Johan Hall, Sandra Kübler, Ryan McDonald, Jens Nilsson, Sebastian Riedel, and Deniz Yuret. 2007. The CoNLL 2007 shared task on dependency parsing. In *Proceedings of EMNLP-CoNLL*.

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2012. A universal part-of-speech tagset. In *Proceedings of LREC*.

Sujith Ravi and Kevin Knight. 2009. Minimized models for unsupervised part-of-speech tagging. In *Proceedings of ACL-IJCNLP*.

11

Main text in numbered sections

Acknowledgments (optional)

References (alphabetical by last name)

The Anatomy of a TACL Style Article

Introduction

- State the research problem and relate it to previous research
- Give a synopsis of the rest of the article

Related work

- Model 1: After introduction, before contributions
- Model 2: After contributions, before conclusion

Contributions

- Theory → Method → Results → Discussion

Conclusion

- Evaluate contributions, point to new research directions

References

- Language technology mostly uses the Harvard system
  - Author-year citations in text
  - Alphabetical list of references at the end (no footnotes)
- Citations in the text:
  - Parenthetical: Parsing is hard (Anderson, 2010).
  - Syntactic: Anderson (2010) claims that parsing is hard.
  - More than two authors:
    - In text, use et al.
      Parsing is hard (Anderson et al., 2010).
      Anderson et al. (2010) claims that parsing is hard.
    - All authors in reference list
      Anderson, P., Svensson, G, Lind, W. and Sund, T. 2017.
      Parsing is hard. . . .

Reference List

- Reference list including all (and only) works cited in the text:
  - Journal article: author, year, title, *journal*, volume, number, pages
  - Conference paper: author, year, title, *proceedings*, pages, location
  - Book chapter: author, year, title, *book*, editors, publisher, pages
  - Book: author, year, *title*, publisher
  - Technical report: author, year, title, organization
  - Thesis: author, year, title, type of thesis, school
- Important: BE CONSISTENT!

Giving Oral Presentations

Preparation is the key

- Think through what you want to say
- Formulate key passages in concrete sentences
- Prepare audiovisual aids (if relevant)

Practice makes perfect

- Rehearse the presentation (many times)
- Time the presentation and note any disfluencies
- Modify and rehearse until fluent

The Structure of Oral Presentations

Oral presentations are basically structured as written reports but

- typically contain less material due to time constraints (especially the background part)
- are often less formal and detailed due to real-time processing (the big picture instead of the formal details)
- can be more repetitive due to memory limitations (get the take-home message across)

The discussion part:

- Listen to the question
- Answer the question – if you can

Audiovisual Aids

Slides provide support for the presentation

- Key points and important concepts
- Graphical illustrations (and sound if relevant)
- Material that is hard to present orally (equations, examples)

But remember

- Not too much information (or too small fontsize) on one slide
- Not running text (to be read aloud)
- Slides should support presentation, not vice versa

Geoff Pullum's Golden Rules

- Don't ever begin with an apology
- Don't ever underestimate the audience's intelligence
- Respect the time limits
- Don't survey the whole damn field
- Remember that you're an advocate, not the defendant
- Expect questions that will floor you