Probabilistic models for sentence level similarity

Daniel Rižnar, Uroš Kosič

University of Ljubljana
Faculty of Computer and Information Science
Tržaška 25, 1000 Ljubljana
{driznar, ukosic}@gmail.com

***Abstract*—Sentence similarity measures become increasingly important in text-related research and application areas such as text mining, information extraction, automatic question-answering, text summarization, text classification and machine translation. In this work, we present an overview of existing methods and models used for this purpose. We implemented two models: probabilistic model with expanded text representation and IBM's translation model 1 with some assumptions. We provide results of experiments with these models. Our measures are based on 65 sentence pairs from Pilot Short Text Semantic Similarity Benchmark Data set. RESULTS!!!**

***Index Terms*—expanded text representation, probabilistic models, sentence similarity, translation models**

I. Introduction

Measuring the similarity between documents and queries has been thoroughly studied in articles related to information retrieval. However, as areas of text mining, information extraction, summarization and machine translation spread, a need of computing the similarity between two very short segments of texts (sentences) has emerged. Unfortunately, standard text similarity measures perform poorly on such tasks because of data sparseness and the lack of context. Such measures rely heavily on terms occurring in both the query and the document. If the query and the document have few terms in common, the score will be low. Because these methods don't use topical relations between the query and the document, they suffer from vocabulary mismatch problem. In sentence level similarity measures the topical relation is very important, but hard to acquire. For example “USA” and “United States of America” are semantically equivalent, yet share no terms in common.

Another problem when measuring the similarity between two sentences, is lack of context information. Standard methods rely on reasonable amount of text in a document, while sentences provide only a limited context. For example, “Apple computer” and “apple pie” share 50% of their terms, but are topically different. Standard methods would rate these two segments as similar, although topically, they are completely different. If we add additional document describing “apple pie”, the results are much better.

In Section 2, we provide an overview of related work, with emphasis on probabilistic models. We examine several approaches to sentence level similarity:

- *Word overlap measures*
  This is a baseline measure, measuring simple overlap of words.
- *TF-IDF measures*
  These measures are a broad class of functions used for estimating relevance and similarity topically between queries and documents.
- *Relative-frequency measures*
  These measures have been shown to perform well at finding co-derivative documents.
- *Probabilistic models*
  Translation transforms text in one language to text in another, with the aim of preserving as much of the semantics as possible. In ___ they propose using statistical translation models in much the same manner to estimate the probability that one sentence is a translation of another.
  In ____ they present probabilistic method with real time expanded representation of sentences. We use this method later in out experimental research.

In Section 3, we present our evaluation experiments and details about data used in measures. We chose data from Pilot Short Text Semantic Similarity Benchmark Data set. Data set contains 65 sentence pairs, already rated by 32 human participants. This data set is widely used as a benchmark for validating short text semantic similarity.

In Section 4, we provide the details of our experimental evaluation of selected methods.

In Section 5, we wrap up and provide conclusions and directions for future work.

II. Related Work

There is a large literature on probabilistic approaches to information retrieval. However, when we want to measure similarity between two sentences, these methods don't perform very well.

In __FLOW___ they focused themselves on intermediate levels of similarity. They explored mechanisms for measuring such intermediate kinds of similarity, focusing on the task of identifying where a particular piece of information originated. Their main topic of research was tracking information flow through various texts. Within their work, they provided some methods for measuring similarity on a sentence level. We will discuss them briefly in following passage.

All their techniques calculate similarity score S(Q, R) between sentences Q and R, intended to capture numerically the extent to which they convey the same information. All methods return maximized S(Q, R), when sentence R has highest degree of similarity to the sentence Q.

A. Word overlap measures

They chose this measure as a baseline measure. Word overlap means the proportion of words in Q that also appear in the candidate sentence R:

$$S(Q,R)=\frac{|Q\cap R|}{|Q|}$$

$|Q\cap R|$ is the number of terms that appear in Q and R
$|Q|$ is the number of all terms that appear in Q

The logic behind word overlap is simple – if two sentences have many terms in common then they are probably similar to some degree.

They also experimented with adjusted version of word overlap, where they took inverse document frequency (IDF) into account. IDF is actually a weight used to evaluate how important a word is to a document in a collection or corpus. The importance increases proportionally to the number of times a word appears in the document but is offset by the frequency of the word in the corpus.

$$S(Q,R)=\frac{|Q\cap R|}{|Q|}\sum_{w\in Q\cap R}\log\left(\frac{N}{df_w}\right)$$

N is total number of documents in the corpus
df_w is number of documents where the term w appears.

Apparently, this adjustment requires additional context (in their terminology additional documents). We will discuss obtaining this additional data later in this section (we call that expended representation of a sentence).

B. TF-IDF measures

Term frequency-inverse document frequency (tf-idf) is actually a weight often used in information retrieval and text mining. This weight is a statistical measure used to evaluate how important a word is to a document in a collection or a corpus. The importance increases proportionally to the number of times a word appears in the document but is offset by the frequency of the word in the corpus. The basic intuitions are that the more frequently a word appears in a passage, the more indicative that word is of the topicality of that passage; and that less frequently a term appears in a collection, the greater its power to discriminate between interesting and uninteresting passages.

$$S(Q,R)=\sum_{w\in Q\cap R}\log(tf_{w,Q}+1)\log(tf_{w,R}+1)\log\left(\frac{N+1}{df_w+0.5}\right)$$

$tf_{w,Q}$ is number of times term *w* appears in sentence *Q*
$tf_{w,R}$ is number of times term *w* appears in sentence *R*
N is the total number of documents in the collection
df_w is the number of documents that *w* appears in

First two summands measure frequency of a term in both sentences, so bigger number means greater participation of a term in both sentences. Last summand is an inverse document frequency, used to lower the score for common, unrepresentative words, such as: the, an, a, be, so...

C. Relative-frequency measures

Relative-frequency measures have been shown to perform well at finding co-derivative documents. In their work, they investigate how well such methods work at finding co-derivative pieces of text at the sentence level.

$$\frac{1}{1+\frac{max(|Q|,|R|)}{min(|Q|,|R|)}}\sum_{w\in Q\cap R}\frac{\log\left(\frac{N}{df_w}\right)}{1+|tf_{w,D}|-|tf_{w,R}|}$$

$max(|Q|,|R|)$ is the length of longer sentence
$min(|Q|,|R|)$ is the length of shorter sentence
Other quantities are defined as above.

The numerator is a standard IDF factor explained above, while the denominator consists of two parts. First part (outside the sum) penalizes differences in the overall lengths of the sentences. Second part (inside the sum) penalizes inequalities in the relative frequency of a word between the two sentences.

D. Probabilistic models

Probabilistic models are based on idea of estimating the probability that one sentence is a translation of another. This translation probability then serves as the basis of the similarity score for pairs of sentences. Statistical machine translation systems aim to generate high-quality translations of sentences between natural languages. Such systems make use of parametrized statistical language models of both source and target language, and a parametrized statistical translation model that estimates the probability that a given target sentence is a translation of the source sentence. Given these models and a parametrization, the system searches a space of possible translations and returns the sentence with the highest probability. In their paper, they propose using statistical translation models in much the same manner to estimate the probability that one sentence is a translation of another. However, as our problem is different from normal translation problems (both sentences are in the same language), we can make some assumptions. We will now briefly summarize their path from more general model to a model adequate to our problem. We will also add some motivation and description of some specific terms.

They start with IBM's Translation Model 1. IBM Model 1 is a generative model. Generative modeling means breaking up the process of generating the data into smaller steps, modeling the smaller steps with probability distributions, and combining the steps into a coherent story. They provide following similarity function, based on IBM model 1:

$$S(Q,R)=\frac{1}{(|R|+1)^{|Q|}}\prod_{i=1}^{|Q|}\sum_{j=1}^{|R|+1}P_t(q_i|r_j)$$

$|R|$ is the length of sentence R

$|Q|$ is the length of sentence Q

$P_t(q_i|r_j)$ is a probability that j-th word in R is a translation of i-th word in q

Then they made some additional assumptions. The original model assumes that each sentence has a special *null* term at position 1; this is the reason that the summation iterates through |R| +1 terms. The null term is used to represent the fact that the current term in Q doesn't align to any terms in R.

With that in mind, they make the distributional assumption that $P_t(q_i|r_1)=P(q_i|C)$, where C is the background model inferred from the collection as a whole. This proceeds from the intuition that - in the absence of any other evidence – an unaligned word is likely to be present in a sentence with a probability equal to its overall probability in the more generalized background language model. The probability of aligning to the null term dictates the influence of the background language model on the resulting translation. Because IBM Model 1 assumes that all reorderings are equally likely, the probability that a term in Q will align to the null term is $\frac{1}{|R|+1}$. Then they generalize the original model by assuming there exists μ null terms in each sentence, where μ is a non-negative integer. This results in each sentence having length |R| + μ, where |R| is the number of non-null terms in R. This model can be described as:

$$S(Q,R)=\frac{1}{(|R|+\mu)^{|Q|}}\prod_{i=1}^{|Q|}\left[\sum_{j=1}^{\mu}P(q_i|C)+\sum_{j=\mu+1}^{|R|+\mu}P_t(q_i|r_j)\right]$$

μ is the number of null terms in each sentence

$P(q_i|C)$ is a probability that i-th term in Q appears in some background model C

They simplify the model further, with assuming that each word translates to itself; that is $P_t(q_i|r_j)=1$ if $q_i=r_j$. This results in the following form:

$$S(Q,R)=\prod_{i=1}^{|Q|}\frac{tf_{q_i,R}+\mu P(q_i|C)}{|R|+\mu}$$

$tf_{q_i,R}$ is the frequency of i-th word in sentence Q in sentence R

Above function is known as language modeling *query likelihood ranking function* using Dirichlet smoothing parameter μ. With μ=1, we get Berger and Lafferty's Translation Model 0. All models here assume that every term only translates to itself. We extended this model with synonyms and so incorporated a more refined estimate of the true translation probabilities. As parameter μ approaches 0, the model becomes word overlap measure that will likely be good at finding exact matches. At the other extreme, as μ gets large more background terms are allowed, which is likely (and known to be) good at finding topically relevant matches.

They defined similarity spectrum, where at one end there is exact identity and at the other general topic relation. They divided this spectrum into 5 parts: exact match, minor edit, same facts, specific topic match, general topic match. They found out that at the general and specific topic level, query likelihood function with μ=2500 gives the best results. This was expected, because past research has shown query likelihood to be effective at identifying topicality. At other levels the relative performance difference between techniques was small, but Translation Model 0 (μ=1) was consistently the most effective.

E. Negative KL-Divergence and expanded representation

In ___SHORTSEGMENTS___ they investigated different similarity measures for short segments of text. They also took different text representation into account:

- surface representation is the most basic representation of a short text segment – the text itself.
- stemmed representation is normalization of a text (ie.

"marine vegetation" becomes "marin veget"). Although stemming can significantly improve matching coverage, it also introduces noise, which can lead to poor matches.

- expanded representation is good for handling contextual problems. If stemming fails to discern the difference between the meaning of "bank" in "Bank of America" and "river bank", expanded representation can gather additional contextual information and perform better in these cases. One approach is to enrich the representation using an external source of information related to query terms. Possible sources of such information include web search results returned by issuing the short text segment as a query, relevant Wikipedia articles... Each of these sources provides a set of contextual text that can be used to expand the original sparse text representation. In their experiments, they used web search results to expand short text representation.

They used this expanded representation with probabilistic measure framework. They define the problem as: given two short segments of text, Q and C, treating Q as the query and C as the candidate we wish to measure similarity of. For ranking purposes they used the negative KL-divergence between query and candidate model.

$$-KL(\Theta_Q,\Theta_C)=H(\Theta_Q)-CE(\Theta_Q,\Theta_C)\equiv\ldots$$

$$\ldots\equiv\sum_{w\in V}P(w|\Theta_Q)\log(P(w|\Theta_C))$$

$H(\Theta_Q)$ is the entropy for the query model

$CE(\Theta_Q,\Theta_C)$ is the cross entropy for query and candidate models

V is the vocabulary (all unique words in both sentences Q and C)

$P(w|\Theta_Q)$ is the estimation of query model

$P(w|\Theta_C)$ is the estimation of candidate model

These estimates are defined as:

$$P(w|\Theta_Q)=\frac{tf_{w,QE}+\mu_Q P(w|C)}{|QE|+\mu_Q}$$

$tf_{w,QE}$ is the frequency of term w in QE

μ_Q is Dirichlet smoothing parameter

$P(w|C)$ is the probability of word w in candidate sentence C

QE is the query sentence Q expanded representation

$$P(w|\Theta_C)=\frac{tf_{w,CE}+\mu_C P(w|C)}{|CE|+\mu_C}$$

$tf_{w,CE}$ is the frequency of term w in CE

CE is the candidate sentence C expanded representation

These estimations are basically the same as query likelihood ranking function presented on previous page.

Their results show that probabilistic methods are good at finding topicality related matches. The probabilistic framework presented in their paper provides a general method for measuring the similarity between two short segments of text.

III. Experimental Evaluation

In this section we present methods and models used in our experiments, data used in our measures and evaluation of these methods.

Ctronic sfile, TRANS-JOUR.DOC, from http://www.ieee.org/organizations/pubs/transactions/stylesheets.htm so you can use it to prepare your manuscript. If you would prefer to use LATEX, download IEEE's LATEX style and sample files from the same Web page. Use these LATEX files for formatting, but please follow the instructions in TRANS-JOUR.DOC or TRANS-JOUR.PDF.

If your paper is intended for a *conference*, please contact your conference editor concerning acceptable word processor formats for your particular conference.

When you open TRANS-JOUR.DOC, select "Page Layout" from the "View" menu in the menu bar (View | Page Layout), which allows you to see the footnotes. Then type over sections of TRANS-JOUR.DOC or cut and paste from another document and then use markup styles. The pull-down style menu is at the left of the Formatting Toolbar at the top of your *Word* window (for example, the style at this point in the document is "Text"). Highlight a section that you want to designate with a certain style, then select the appropriate name on the style menu. The style will adjust your fonts and line spacing. **Do not change the font sizes or line spacing to squeeze more text into a limited number of pages.** Use italics for emphasis; do not underline.

To insert images in *Word*, position the cursor at the insertion point and either use Insert | Picture | From File or copy the image to the Windows clipboard and then Edit | Paste Special | Picture (with "Float over text" unchecked).

IEEE will do the final formatting of your paper. If your paper is intended for a conference, please observe the conference page limits.

IV. Procedure for Paper Submission

A. Review Stage

Please check with your editor on whether to submit your manuscript by hard copy or electronically for review. If hard copy, submit photocopies such that only one column appears per page. This will give your referees plenty of room to write comments. Send the number of copies specified by your editor (typically four). If submitted electronically, find out if your editor prefers submissions on disk or as e-mail attachments.

If you want to submit your file with one column

electronically, please do the following:

--First, click on the View menu and choose Print Layout.

--Second, place your cursor in the first paragraph. Go to the Format menu, choose Columns, choose one column Layout, and choose "apply to whole document" from the dropdown menu.

--Third, click and drag the right margin bar to just over 4 inches in width.

The graphics will stay in the "second" column, but you can drag them to the first column. Make the graphic wider to push out any text that may try to fill in next to the graphic.

B. Final Stage

When you submit your final version, after your paper has been accepted, print it in two-column format, including figures and tables. Send three prints of the paper; two will go to IEEE and one will be retained by the Editor-in-Chief or conference publications chair.

You must also send your final manuscript on a disk, which IEEE will use to prepare your paper for publication. Write the authors' names on the disk label. If you are using a Macintosh, please save your file on a PC formatted disk, if possible. You may use *Zip* or CD-ROM disks for large files, or compress files using *Compress*, *Pkzip*, *Stuffit*, or *Gzip*.

Also send a sheet of paper with complete contact information for all authors. Include full mailing addresses, telephone numbers, fax numbers, and e-mail addresses. This information will be used to send each author a complimentary copy of the journal in which the paper appears. In addition, designate one author as the "corresponding author." This is the author to whom proofs of the paper will be sent. Proofs are sent to the corresponding author only.

C. Figures

All tables and figures will be processed as images. **However, IEEE cannot extract the tables and figures embedded in your document.** (The figures and tables you insert in your document are only to help you gauge the size of your paper, for the convenience of the referees, and to make it easy for you to distribute preprints.) Therefore, **submit, on separate sheets of paper, enlarged versions of the tables and figures that appear in your document.** These are the images IEEE will scan and publish with your paper.

D. Electronic Image Files (Optional)

You will have the greatest control over the appearance of your figures if you are able to prepare electronic image files. If you do not have the required computer skills, just submit paper prints as described above and skip this section.

1) Easiest Way: If you have a scanner, the best and quickest way to prepare noncolor figure files is to print your tables and figures on paper exactly as you want them to appear, scan them, and then save them to a file in PostScript (PS) or Encapsulated PostScript (EPS) formats. Use a separate file for each image. File names should be of the form "fig1.ps" or "fig2.eps."

2) Slightly Harder Way: Using a scanner as above, save the images in TIFF format. High-contrast line figures and tables should be prepared with 600 dpi resolution and saved with no compression, 1 bit per pixel (monochrome), with file names of the form "fig3.tif" or "table1.tif." To obtain a 3.45-in figure (one-column width) at 600 dpi, the figure requires a horizontal size of 2070 pixels. Typical file sizes will be on the order of 0.5 MB.

Photographs and grayscale figures should be prepared with 220 dpi resolution and saved with no compression, 8 bits per pixel (grayscale). To obtain a 3.45-in figure (one-column width) at 220 dpi, the figure should have a horizontal size of 759 pixels.

Color figures should be prepared with 400 dpi resolution and saved with no compression, 8 bits per pixel (palette or 256 color). To obtain a 3.45-in figure (one column width) at 400 dpi, the figure should have a horizontal size of 1380 pixels.

For more information on TIFF files, please go to http://www.ieee.org/organizations/pubs/transactions/information.htm and click on the link "Guidelines for Author Supplied Electronic Text and Graphics."

3) Somewhat Harder Way: If you do not have a scanner, you may create noncolor PostScript figures by "printing" them to files. First, download a PostScript printer driver from http://www.adobe.com/support/downloads/pdrvwin.htm (for Windows) or from http://www.adobe.com/support/downloads/ pdrvmac.htm (for Macintosh) and install the "Generic PostScript Printer" definition. In *Word*, paste your figure into a new document. Print to a file using the PostScript printer driver. File names should be of the form "fig5.ps." Use Adobe Type 1 fonts when creating your figures, if possible.

4) Other Ways: Experienced computer users can convert figures and tables from their original format to TIFF. Some useful image converters are Adobe *Photoshop*, Corel *Draw*, and Microsoft *Photo Editor*, an application that is part of Microsoft *Office 97* and *Office 2000* (look for C:\Program Files\Common Files \Microsoft Shared\ PhotoEd\ PHOTOED.EXE. (You may have to custom-install *Photo Editor* from your original *Office* disk.)

Here is a way to make TIFF image files of tables. First, create your table in *Word*. Use horizontal lines but no vertical lines. Hide gridlines (Table | Hide Gridlines). Spell check the table to remove any red underlines that indicate spelling errors. Adjust magnification (View | Zoom) such that you can view the entire table *at maximum area* when you select View | Full Screen. Move the cursor so that it is out of the way. Press "Print Screen" on your keyboard; this copies the screen image to the Windows clipboard. Open Microsoft *Photo Editor* and click Edit | Paste as New Image. Crop the table image (click Select button; select the part you want, then Image | Crop). Adjust the properties of the image (File | Properties) to monochrome (1 bit) and 600 pixels per inch. Resize the image (Image | Resize) to a width of 3.45 inches. Save the file (File | Save As) in TIFF with no compression (click "More" button).

Most graphing programs allow you to save graphs in TIFF; however, you often have no control over compression or number of bits per pixel. You should open these image files

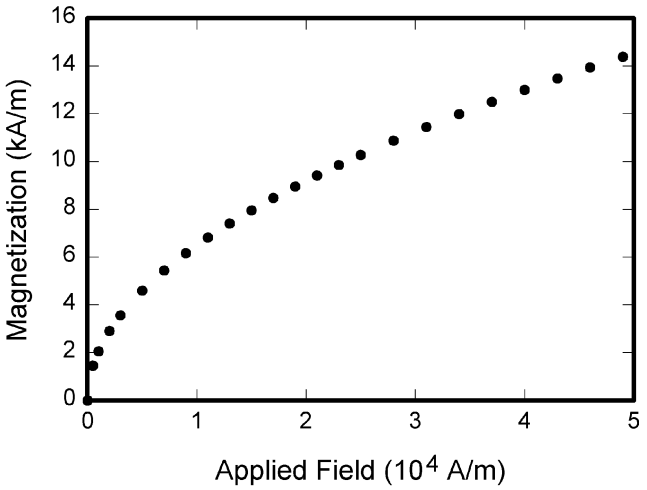

Fig. 1. Magnetization as a function of applied field. Note that "Fig." is abbreviated. There is a period after the figure number, followed by two spaces. It is good practice to explain the significance of the figure in the caption.

in a program such as Microsoft *Photo Editor* and re-save them using no compression, either 1 or 8 bits, and either 600 or 220 dpi resolution (File | Properties; Image | Resize). See Section II-D2 for an explanation of number of bits and resolution. If your graphing program cannot export to TIFF, you can use the same technique described for tables in the previous paragraph.

A way to convert a figure from Windows Metafile (WMF) to TIFF is to paste it into Microsoft *PowerPoint*, save it in JPG format, open it with Microsoft *Photo Editor* or similar converter, and re-save it as TIFF.

Microsoft *Excel* allows you to save spreadsheet charts in Graphics Interchange Format (GIF). To get good resolution, make the *Excel* charts *very* large. Then use the "Save as HTML" feature (see http://support.microsoft.com/support/kb/articles/q158/0/79.asp). You can then convert from GIF to TIFF using Microsoft *Photo Editor*, for example.

No matter how you convert your images, it is a good idea to print the TIFF files to make sure nothing was lost in the conversion.

If you modify this document for use with other IEEE journals or conferences, you should save it as type "Word 97-2000 & 6.0/95 - RTF (*.doc)" so that it can be opened by any version of *Word.*

TABLE I
UNITS FOR MAGNETIC PROPERTIES

Symbol	Quantity	Conversion from Gaussian and CGS EMU to SI [a]
Φ	magnetic flux	1 Mx $\rightarrow 10^{-8}$ Wb = 10^{-8} V·s
B	magnetic flux density, magnetic induction	1 G $\rightarrow 10^{-4}$ T = 10^{-4} Wb/m^2
H	magnetic field strength	1 Oe $\rightarrow 10^3/(4\pi)$ A/m
m	magnetic moment	1 erg/G = 1 emu $\rightarrow 10^{-3}$ A·m^2 = 10^{-3} J/T
M	magnetization	1 erg/(G·cm^3) = 1 emu/cm^3 $\rightarrow 10^3$ A/m
$4\pi M$	magnetization	1 G $\rightarrow 10^3/(4\pi)$ A/m
σ	specific magnetization	1 erg/(G·g) = 1 emu/g $\rightarrow$ 1 A·m^2/kg
j	magnetic dipole moment	1 erg/G = 1 emu $\rightarrow 4\pi \times 10^{-10}$ Wb·m
J	magnetic polarization	1 erg/(G·cm^3) = 1 emu/cm^3 $\rightarrow 4\pi \times 10^{-4}$ T
χ, κ	susceptibility	1 $\rightarrow 4\pi$
χ_ρ	mass susceptibility	1 cm^3/g $\rightarrow 4\pi \times 10^{-3}$ m^3/kg
μ	permeability	1 $\rightarrow 4\pi \times 10^{-7}$ H/m = $4\pi \times 10^{-7}$ Wb/(A·m)
μ_r	relative permeability	$\mu \rightarrow \mu_r$
w, W	energy density	1 erg/cm^3 $\rightarrow 10^{-1}$ J/m^3
N, D	demagnetizing factor	1 $\rightarrow 1/(4\pi)$

No vertical lines in table. Statements that serve as captions for the entire table do not need footnote letters.

[a]Gaussian units are the same as cgs emu for magnetostatics; Mx = maxwell, G = gauss, Oe = oersted; Wb = weber, V = volt, s = second, T = tesla, m = meter, A = ampere, J = joule, kg = kilogram, H = henry.

E. Copyright Form

An IEEE copyright form should accompany your final submission. You can get a .pdf, .html, or .doc version at http://www.ieee.org/copyright or from the first issues in each volume of the IEEE TRANSACTIONS and JOURNALS. Authors are responsible for obtaining any security clearances.

V. MATH

If you are using *Word*, use either the Microsoft Equation Editor or the *MathType* add-on (http://www.mathtype.com) for equations in your paper (Insert | Object | Create New | Microsoft Equation *or* MathType Equation). "Float over text" should *not* be selected.

VI. UNITS

Use either SI (MKS) or CGS as primary units. (SI units are strongly encouraged.) English units may be used as secondary units (in parentheses). **This applies to papers in data storage.** For example, write "15 Gb/cm^2 (100 Gb/in^2)." An exception is when English units are used as identifiers in trade, such as "3½ in disk drive." Avoid combining SI and CGS units, such as current in amperes and magnetic field in oersteds. This often leads to confusion because equations do not balance dimensionally. If you must use mixed units, clearly state the units for each quantity in an equation.

The SI unit for magnetic field strength H is A/m. However, if you wish to use units of T, either refer to magnetic flux density B or magnetic field strength symbolized as $\mu_0 H$. Use the center dot to separate compound units, e.g., "A·m^2."

VII. Helpful Hints

A. Figures and Tables

Because IEEE will do the final formatting of your paper, you do not need to position figures and tables at the top and bottom of each column. In fact, all figures, figure captions, and tables can be at the end of the paper. Large figures and tables may span both columns. Place figure captions below the figures; place table titles above the tables. If your figure has two parts, include the labels "(a)" and "(b)" as part of the artwork. Please verify that the figures and tables you mention in the text actually exist. **Please do not include captions as part of the figures. Do not put captions in "text boxes" linked to the figures. Do not put borders around the outside of your figures.** Use the abbreviation "Fig." even at the beginning of a sentence. Do not abbreviate "Table." Tables are numbered with Roman numerals.

Color printing of figures is available, but is billed to the authors (approximately $1300, depending on the number of figures and number of pages containing color). Include a note with your final paper indicating that you request color printing. **Do not use color unless it is necessary for the proper interpretation of your figures.** If you want reprints of your color article, the reprint order should be submitted promptly. There is an additional charge of $81 per 100 for color reprints.

Figure axis labels are often a source of confusion. Use words rather than symbols. As an example, write the quantity "Magnetization," or "Magnetization *M*," not just "*M*." Put units in parentheses. Do not label axes only with units. As in Fig. 1, for example, write "Magnetization (A/m)" or "Magnetization (A m^{-1})," not just "A/m." Do not label axes with a ratio of quantities and units. For example, write "Temperature (K)," not "Temperature/K."

Multipliers can be especially confusing. Write "Magnetization (kA/m)" or "Magnetization (10^3 A/m)." Do not write "Magnetization (A/m) × 1000" because the reader would not know whether the top axis label in Fig. 1 meant 16000 A/m or 0.016 A/m. Figure labels should be legible, approximately 8 to 12 point type.

B. References

Number citations consecutively in square brackets [1]. The sentence punctuation follows the brackets [2]. Multiple references [2], [3] are each numbered with separate brackets [1]–[3]. When citing a section in a book, please give the relevant page numbers [2]. In sentences, refer simply to the reference number, as in [3]. Do not use "Ref. [3]" or "reference [3]" except at the beginning of a sentence: "Reference [3] shows ... ." Unfortunately the IEEE document translator cannot handle automatic endnotes in *Word*; therefore, type the reference list at the end of the paper using the "References" style.

Number footnotes separately in superscripts (Insert | Footnote).[1] Place the actual footnote at the bottom of the column in which it is cited; do not put footnotes in the reference list (endnotes). Use letters for table footnotes (see Table I).

Please note that the references at the end of this document are in the preferred referencing style. Give all authors' names; do not use "*et al*." unless there are six authors or more. Use a space after authors' initials. Papers that have not been published should be cited as "unpublished" [4]. Papers that have been submitted for publication should be cited as "submitted for publication" [5]. Papers that have been accepted for publication, but not yet specified for an issue should be cited as "to be published" [6]. Please give affiliations and addresses for private communications [7].

Capitalize only the first word in a paper title, except for proper nouns and element symbols. For papers published in translation journals, please give the English citation first, followed by the original foreign-language citation [8].

C. Abbreviations and Acronyms

Define abbreviations and acronyms the first time they are used in the text, even after they have already been defined in the abstract. Abbreviations such as IEEE, SI, ac, and dc do not have to be defined. Abbreviations that incorporate periods should not have spaces: write "C.N.R.S.," not "C. N. R. S." Do not use abbreviations in the title unless they are unavoidable (for example, "IEEE" in the title of this article).

D. Equations

Number equations consecutively with equation numbers in parentheses flush with the right margin, as in (1). First use the equation editor to create the equation. Then select the "Equation" markup style. Press the tab key and write the equation number in parentheses. To make your equations more compact, you may use the solidus (/), the exp function, or appropriate exponents. Use parentheses to avoid ambiguities in denominators. Punctuate equations when they are part of a sentence, as in

$$\int F(r,\phi)\, dr\, d\phi = [\sigma r_2 (2\mu_0)] \cdot \int \exp(-\lambda |z_j - z_i|)\, \lambda^{-1} J_1(\lambda r_2)\, J_0(\lambda r_i)\, d\lambda . \quad (1)$$

Be sure that the symbols in your equation have been defined before the equation appears or immediately following. Italicize symbols (*T* might refer to temperature, but T is the unit tesla). Refer to "(1)," not "Eq. (1)" or "equation (1)," except at the beginning of a sentence: "Equation (1) is ... ."

E. Other Recommendations

Use one space after periods and colons. Hyphenate complex modifiers: "zero-field-cooled magnetization." Avoid dangling participles, such as, "Using (1), the potential was calculated." [It is not clear who or what used (1).] Write instead, "The potential was calculated by using (1)," or "Using (1), we calculated the potential."

Use a zero before decimal points: "0.25," not ".25." Use

[1] It is recommended that footnotes be avoided (except for the unnumbered footnote with the receipt date on the first page). Instead, try to integrate the footnote information into the text.

"cm^3," not "cc." Indicate sample dimensions as "0.1 cm × 0.2 cm," not "0.1 × 0.2 cm^2." The abbreviation for "seconds" is "s," not "sec." Do not mix complete spellings and abbreviations of units: use "Wb/m^2" or "webers per square meter," not "webers/m^2." When expressing a range of values, write "7 to 9" or "7-9," not "7~9."

A parenthetical statement at the end of a sentence is punctuated outside of the closing parenthesis (like this). (A parenthetical sentence is punctuated within the parentheses.) In American English, periods and commas are within quotation marks, like "this period." Other punctuation is "outside"! Avoid contractions; for example, write "do not" instead of "don't." The serial comma is preferred: "A, B, and C" instead of "A, B and C."

If you wish, you may write in the first person singular or plural and use the active voice ("I observed that ..." or "We observed that ..." instead of "It was observed that ..."). Remember to check spelling. If your native language is not English, please get a native English-speaking colleague to proofread your paper.

VIII. Some Common Mistakes

The word "data" is plural, not singular. The subscript for the permeability of vacuum μ_0 is zero, not a lowercase letter "o." The term for residual magnetization is "remanence"; the adjective is "remanent"; do not write "remnance" or "remnant." Use the word "micrometer" instead of "micron." A graph within a graph is an "inset," not an "insert." The word "alternatively" is preferred to the word "alternately" (unless you really mean something that alternates). Use the word "whereas" instead of "while" (unless you are referring to simultaneous events). Do not use the word "essentially" to mean "approximately" or "effectively." Do not use the word "issue" as a euphemism for "problem." When compositions are not specified, separate chemical symbols by en-dashes; for example, "NiMn" indicates the intermetallic compound $Ni_{0.5}Mn_{0.5}$ whereas "Ni–Mn" indicates an alloy of some composition Ni_xMn_{1-x}.

Be aware of the different meanings of the homophones "affect" (usually a verb) and "effect" (usually a noun), "complement" and "compliment," "discreet" and "discrete," "principal" (e.g., "principal investigator") and "principle" (e.g., "principle of measurement"). Do not confuse "imply" and "infer."

Prefixes such as "non," "sub," "micro," "multi," and ""ultra" are not independent words; they should be joined to the words they modify, usually without a hyphen. There is no period after the "et" in the Latin abbreviation "*et al.*" (it is also italicized). The abbreviation "i.e.," means "that is," and the abbreviation "e.g.," means "for example" (these abbreviations are not italicized).

An excellent style manual and source of information for science writers is [9]. A general IEEE style guide, *Information for Authors*, is available at http://www.ieee.org/organizations/pubs/transactions/information.htm

IX. Editorial Policy

Submission of a manuscript is not required for participation in a conference. Do not submit a reworked version of a paper you have submitted or published elsewhere. Do not publish "preliminary" data or results. The submitting author is responsible for obtaining agreement of all coauthors and any consent required from sponsors before submitting a paper. IEEE TRANSACTIONS and JOURNALS strongly discourage courtesy authorship. It is the obligation of the authors to cite relevant prior work.

The Transactions and Journals Department does not publish conference records or proceedings. The TRANSACTIONS does publish papers related to conferences that have been recommended for publication on the basis of peer review. As a matter of convenience and service to the technical community, these topical papers are collected and published in one issue of the TRANSACTIONS.

At least two reviews are required for every paper submitted. For conference-related papers, the decision to accept or reject a paper is made by the conference editors and publications committee; the recommendations of the referees are advisory only. Undecipherable English is a valid reason for rejection. Authors of rejected papers may revise and resubmit them to the TRANSACTIONS as regular papers, whereupon they will be reviewed by two new referees.

X. Publication Principles

The contents of IEEE TRANSACTIONS and JOURNALS are peer-reviewed and archival. The TRANSACTIONS publishes scholarly articles of archival value as well as tutorial expositions and critical reviews of classical subjects and topics of current interest.

Authors should consider the following points:

1) Technical papers submitted for publication must advance the state of knowledge and must cite relevant prior work.
2) The length of a submitted paper should be commensurate with the importance, or appropriate to the complexity, of the work. For example, an obvious extension of previously published work might not be appropriate for publication or might be adequately treated in just a few pages.
3) Authors must convince both peer reviewers and the editors of the scientific and technical merit of a paper; the standards of proof are higher when extraordinary or unexpected results are reported.
4) Because replication is required for scientific progress, papers submitted for publication must provide sufficient information to allow readers to perform similar experiments or calculations and use the reported results. Although not everything need be disclosed, a paper must contain new, useable, and fully described information. For example, a specimen's chemical composition need not be reported if the main purpose of a paper is to introduce a new measurement technique. Authors should expect to be challenged by reviewers if the results are not supported by adequate data and critical details.
5) Papers that describe ongoing work or announce the latest technical achievement, which are suitable for

presentation at a professional conference, may not be appropriate for publication in a TRANSACTIONS or JOURNAL.

XI. CONCLUSION

A conclusion section is not required. Although a conclusion may review the main points of the paper, do not replicate the abstract as the conclusion. A conclusion might elaborate on the importance of the work or suggest applications and extensions.

APPENDIX

Appendixes, if needed, appear before the acknowledgment.

ACKNOWLEDGMENT

The preferred spelling of the word "acknowledgment" in American English is without an "e" after the "g." Use the singular heading even if you have many acknowledgments. Avoid expressions such as "One of us (S.B.A.) would like to thank ... ." Instead, write "F. A. Author thanks ... ." **Sponsor and financial support acknowledgments are placed in the unnumbered footnote on the first page**.

REFERENCES

[1] G. O. Young, "Synthetic structure of industrial plastics (Book style with paper title and editor)," in *Plastics*, 2nd ed. vol. 3, J. Peters, Ed. New York: McGraw-Hill, 1964, pp. 15–64.

[2] W.-K. Chen, *Linear Networks and Systems* (Book style). Belmont, CA: Wadsworth, 1993, pp. 123–135.

[3] H. Poor, *An Introduction to Signal Detection and Estimation*. New York: Springer-Verlag, 1985, ch. 4.

[4] B. Smith, "An approach to graphs of linear forms (Unpublished work style)," unpublished.

[5] E. H. Miller, "A note on reflector arrays (Periodical style—Accepted for publication)," *IEEE Trans. Antennas Propagat.*, to be published.

[6] J. Wang, "Fundamentals of erbium-doped fiber amplifiers arrays (Periodical style—Submitted for publication)," *IEEE J. Quantum Electron.*, submitted for publication.

[7] C. J. Kaufman, Rocky Mountain Research Lab., Boulder, CO, private communication, May 1995.

[8] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on magneto-optical media and plastic substrate interfaces(Translation Journals style)," *IEEE Transl. J. Magn.Jpn.*, vol. 2, Aug. 1987, pp. 740–741 [*Dig. 9th Annu. Conf. Magnetics* Japan, 1982, p. 301].

[9] M. Young, *The Techincal Writers Handbook*. Mill Valley, CA: University Science, 1989.

[10] J. U. Duncombe, "Infrared navigation—Part I: An assessment of feasibility (Periodical style)," *IEEE Trans. Electron Devices*, vol. ED-11, pp. 34–39, Jan. 1959.

[11] S. Chen, B. Mulgrew, and P. M. Grant, "A clustering technique for digital communications channel equalization using radial basis function networks," *IEEE Trans. Neural Networks*, vol. 4, pp. 570–578, July 1993.

[12] R. W. Lucky, "Automatic equalization for digital communication," *Bell Syst. Tech. J.*, vol. 44, no. 4, pp. 547–588, Apr. 1965.

[13] S. P. Bingulac, "On the compatibility of adaptive controllers (Published Conference Proceedings style)," in *Proc. 4th Annu. Allerton Conf. Circuits and Systems Theory*, New York, 1994, pp. 8–16.

[14] G. R. Faulhaber, "Design of service systems with priority reservation," in *Conf. Rec. 1995 IEEE Int. Conf. Communications*, pp. 3–8.

[15] W. D. Doyle, "Magnetization reversal in films with biaxial anisotropy," in *1987 Proc. INTERMAG Conf.*, pp. 2.2-1–2.2-6.

[16] G. W. Juette and L. E. Zeffanella, "Radio noise currents n short sections on bundle conductors (Presented Conference Paper style)," presented at the IEEE Summer power Meeting, Dallas, TX, June 22–27, 1990, Paper 90 SM 690-0 PWRS.

[17] J. G. Kreifeldt, "An analysis of surface-detected EMG as an amplitude-modulated noise," presented at the 1989 Int. Conf. Medicine and Biological Engineering, Chicago, IL.

[18] J. Williams, "Narrow-band analyzer (Thesis or Dissertation style)," Ph.D. dissertation, Dept. Elect. Eng., Harvard Univ., Cambridge, MA, 1993.

[19] N. Kawasaki, "Parametric study of thermal and chemical nonequilibrium nozzle flow," M.S. thesis, Dept. Electron. Eng., Osaka Univ., Osaka, Japan, 1993.

[20] J. P. Wilkinson, "Nonlinear resonant circuit devices (Patent style)," U.S. Patent 3 624 12, July 16, 1990.

[21] *IEEE Criteria for Class IE Electric Systems* (Standards style), IEEE Standard 308, 1969.

[22] *Letter Symbols for Quantities*, ANSI Standard Y10.5-1968.

[23] R. E. Haskell and C. T. Case, "Transient signal propagation in lossless isotropic plasmas (Report style)," USAF Cambridge Res. Lab., Cambridge, MA Rep. ARCRL-66-234 (II), 1994, vol. 2.

[24] E. E. Reber, R. L. Michell, and C. J. Carter, "Oxygen absorption in the Earth's atmosphere," Aerospace Corp., Los Angeles, CA, Tech. Rep. TR-0200 (420-46)-3, Nov. 1988.

[25] (Handbook style) *Transmission Systems for Communications*, 3rd ed., Western Electric Co., Winston-Salem, NC, 1985, pp. 44–60.

[26] *Motorola Semiconductor Data Manual*, Motorola Semiconductor Products Inc., Phoenix, AZ, 1989.

[27] (Basic Book/Monograph Online Sources) J. K. Author. (year, month, day). *Title* (edition) [Type of medium]. Volume(issue). Available: http://www.(URL)

[28] J. Jones. (1991, May 10). Networks (2nd ed.) [Online]. Available: http://www.atm.com

[29] (Journal Online Sources style) K. Author. (year, month). Title. *Journal* [Type of medium]. Volume(issue), paging if given. Available: http://www.(URL)

[30] R. J. Vidmar. (1992, August). On the use of atmospheric plasmas as electromagnetic reflectors. *IEEE Trans. Plasma Sci.* [Online]. *21(3).* pp. 876—880. Available: http://www.halcyon.com/pub/journals/21ps03-vidmar

First A. Author (M'76–SM'81–F'87) and the other authors may include biographies at the end of regular papers. Biographies are often not included in conference-related papers. This author became a Member (M) of IEEE in 1976, a Senior Member (SM) in 1981, and a Fellow (F) in 1987. The first paragraph may contain a place and/or date of birth (list place, then date). Next, the author's educational background is listed. The degrees should be listed with type of degree in what field, which institution, city, state or country, and year degree was earned. The author's major field of study should be lower-cased.

The second paragraph uses the pronoun of the person (he or she) and not the author's last name. It lists military and work experience, including summer and fellowship jobs. Job titles are capitalized. The current job must have a location; previous positions may be listed without one. Information concerning previous publications may be included. Try not to list more than three books or published articles. The format for listing publishers of a book within the biography is: title of book (city, state: publisher name, year) similar to a reference. Current and previous research interests ends the paragraph.

The third paragraph begins with the author's title and last name (e.g., Dr. Smith, Prof. Jones, Mr. Kajor, Ms. Hunter). List any memberships in professional societies other than the IEEE. Finally, list any awards and work for IEEE committees and publications. If a photograph is provided, the biography will be indented around it. The photograph is placed at the top left of the biography. Personal hobbies will be deleted from the biography.